古文書文字認識システムの高精度化に関する研究

AUTHOR(S):

柴山, 守

古文書文字認識システムの高精度化に関する研究

課題番号　14380184

平成 14 年度～平成 16 年度科学研究費補助金基盤研究(B)(1)研究成果報告書

平成 17 年 5 月

研究代表者　　柴　山　　守

(京都大学東南アジア研究所・教授)

はしがき

HCR(Historical Character Recognition)プロジェクトは、平成11年度の開始からすでに5年が経とうとしている。初期の研究を支えた4つの科学研究費補助金（平成11～13年度基盤研究(B)(1)「古文書解読プロセスの知能情報学的解明」、同「古文書OCRの試論的研究」、同「手書き文字OCR技術を援用した古文書翻刻支援システムの開発」、平成12～14年度基盤研究(B)(1)「古文書解読支援システムの開発と電子辞書技術の応用に関する研究」）が一昨年度までに終了し、プロジェクトは第1期から第2期へと入りつつある。この報告書は、HCRプロジェクトの第1期成果と、第2期での取り組み、今後の課題について報告するものである。

本研究は、日本語手書き文字認識技術を発展的に応用して、古文書OCR機能を盛り込んだ古文書翻刻支援のためのシステムを開発するという、大胆な目論見のもとに進められてきた。過去にあまり例のない研究であることから、研究について関心をお持ちの方もあることだろう。

HCRプロジェクトは、古文書文字データベースの作成、古文書文字の切り出しと認識手法の研究、知識による翻刻支援、電子化古文書文字辞典の開発などにおいて具体的な成果を挙げることができた。これまでの成果について、みなさまからの忌憚のないご意見を頂戴できれば幸いである。

本報告書は、日本学術振興会科学研究費補助金の平成14～16年度基盤研究(B)(1)「古文書文字認識システムの高精度化に関する研究」（課題番号14380184）の研究成果報告書として刊行するものである。日頃から当研究課題にご支援くださっている方々に、あらためて謝意を表したい。

研究代表者　柴　山　　守

Ⅰ．研究組織

研究代表者：柴山　守（京都大学・東南アジア研究所・教授）

研究分担者：加藤　寧（東北大学・大学院情報科学研究科・教授）

山田奨治（国際日本文化研究センター・研究部・助教授）

並木美太郎（東京農工大学・工学部・助教授）

小島正美（東北工業大学・工学部・教授）

梅田三千雄（大阪電気通信大学・総合情報学部・教授）

原　正一郎（国文学研究資料館・研究情報部・助教授）

川口　洋（帝塚山大学・経営情報学部・教授）

石谷　康人（(株) 東芝・研究開発センター・研究主務）

交付決定額（配分額）　　（金額単位：千円）

	直接経費	間接経費	合計
平成 14 年度	2,900	0	2,900
平成 15 年度	4,700	0	4,700
平成 16 年度	4,700	0	4,700
総計	12,300	0	12,300

Ⅱ. 研究発表

HCR プロジェクトのホームページは，http://www.nichibun.ac.jp/shoji/hcr/ である．
最新の研究成果報告や本報告で述べた成果物の公開は，当ホームページからおこなっている．

(1)学会誌等

[1]山田奨治, 柴山　守：古文書を対象にした文字認識の研究, 情報処理, Vol.43, No.9, pp.950－955, 平成14年 9月

[2]梅田三千雄、橋本智広：認識処理を援用した文字切り出しによる古文書キャラクタスポッティング、電気学会論文誌、Vol.122, No.11, pp.1876-1884, 2002

[3]川口　洋：『江戸時代における人口分析システム（Danjuro Ver.2.0)』の構築・運用・利用、帝塚山大学学術論集、No.9, pp.1-27, 2002.12

[4]安倍広多, 中塚麻記子, 柴山　守：『くずし字解読辞典』文字画像からの筆順抽出の試み, 大阪市立大学学術情報総合センター紀要, Vol.4, 平成15年 3月

[5]和泉勇治、海老澤則之、加藤　寧、根本義章：非線形正規化を応用した学習パターン生成による手書き文字認識、電子情報通信学会論文誌、J86-D-II, 10、pp.1391-1399, 2003

[6]H. Nakayama, Y.Waizumi, Nei. Kato, Mamoru. Shibayama, A Nonlinear Shape Normalization Method for Holistic Recognition of Japanese Historical String, Journal of International Journal on Document Analysis and Recognition (forthcoming), 2005

(2)口頭発表

[1]山田奨治, 和泉勇治, 加藤　寧, 柴山　守：類似文字検索機能をそなえた電子くずし字辞典の開発, 情報処理学会研究報告 2002-CH-54, Vol.2002, No.23, pp.43-50, 平成14年 5月

[2]原正一郎：古文書 OCR のための文字切り出し, 情報処理学会研究報告 2002-CH-55, Vol.2002, No.3, pp.43-50, 平成14年 7月

[3]近藤博人, 松本隆一, 柴山　守, 山田奨治, 荒木義：文字切出しを前提としない古文書標題認識, 情報処理学会研究報告 2003-CH-57, Vol.2003, No.5, pp.1-8, 平成15年 1月

[4]篠原早苗、和泉勇治、加藤　寧、根本義章：SVM を用いた手書き文字認識における学習データ選択と認識精度に関する一考察、電子情報通信学会技術研究報告, Vol.102, No.708 PRMU2002-256, pp.81-86, 2003

[5]*Digital Archives using XML Description and Application to Historical Resources*, Proceedings of the Sixth REKIHAKU International Symposium, pp.31－38, 平成15年 2月

[6]証文類古文書標題認識と辞書構築、東洋学へのコンピュータ利用第14回研究セミナー、京都大学人文科学研究所附属漢字情報研究センター・京都大学学術情報メディアセンター、平成15年 3月

[7]山田奨治、柴山　守：n-gram と OCR による定型表現がある古文書の文字の推定、情報処理学会研究報告 2003-CH-58, Vol.2003, No.12, pp.17-22、平成 15 年 5 月

[8]松本隆一、増田好克、柴山　守、荒木義彦：古文書における Hough 変換を用いた行抽出手法の提案、平成 16 年度電気学会全国大会講演論文集、p.109、平成 16 年 3

(3)出版物

[1]古文書文字データベース（HCD）　Web サイトからダウンロード可能

▼HCD1

「宗門改帳」から採字した年齢表記文字 16 字種「ツ」「一」「二」「三」「四」「五」「六」「七」「八」「九」「十」「壱」「弐」「年」「拾」「廿」計 3,066 文字の 2 値画像。川口洋氏作成

Win 版(410KB)　Unix 版(460KB)

▼HCD1a

「宗門改帳」から採字した単位表記文字 16 字種「田」「畑」「高」「石」「斗」「升」「合」「金」「両」「分」「朱」「家」「軒」「間」「馬」「疋」計 3,200 文字の 2 値画像。川口洋氏作成

Win 版(610KB)　Unix 版(680KB)

▼HCD1b

「宗門改帳」から採字した単位表記文字 8 字種「内」「人」「男」「女」「〆」「長」「横」「夕」計 1,600 文字の 2 値画像。川口洋氏作成

Win 版(270KB)　Unix 版(250KB)

▼HCD1c

「宗門改帳」から採字した親族関係表記文字 8 字種「父」「母」「子」「悴」「祖」「弟」「娘」「房」計 1,600 文字の 2 値画像。川口洋氏作成

Win 版(360KB)

▼HCD2

古文書文字切り出し研究用データベース。大阪市立大学所蔵「伏見屋善兵衛文書」（金子借用証文類）から採取した 200 標題行 1,378 文字の 2 値画像

ダウンロード(1MB)

▼HCD2a

古文書文字切り出し研究用データベース。200 標題行。HCD2 の白黒階調画像版

ダウンロード(8.3MB)

▼HCD2b

古文書文字切り出し研究用データベース。200 標題行。HCD2 のフルカラー画像版

ダウンロード(6.8MB)

▼HCD3

「伏見屋善兵衛文書」の 900 標題から切り出した 184 字種 4,933 文字の 2 値画像

ダウンロード(3MB)

[2]HCR ソフトウェア　　Web サイトからダウンロード可能

▼GetAMoji マクロ

古文書翻刻中に遭遇する不明文字（ゲタ文字）の正解候補を提示する機能をもった Microsoft Word のためのマクロ。n-gram を利用。辞書作成機能付き

LZH 形式(480KB)　自己解凍形式(505KB)　関連論文

▼Web 版　GetAMoji

古文書翻刻中に遭遇する不明文字（ゲタ文字）の正解候補を提示する機能をもった GetAMoji マクロの Web 版　Web-GetAMoji へのリンク　関連論文

目次

第Ⅰ部

本文編

第1章

プロジェクトの概況

1.1 問題の所在

われわれは，日本語手書き文字認識を発展的応用する研究として，古文書を対象にした文字認識の研究，およびそれを可能にするための環境の整備，既存の技術を活用した古文書の翻刻（古文書を読んで活字にすること）支援のシステム化の研究などに取り組んでいる．

古文書とは，狭い意味では差出人がある意思伝達の書類のことであるが，この報告でいう古文書は，他者への意思伝達を目的としない「古記録」や「古典籍」も含めた，広い意味で捉えることにする．時代でいうならば，古代から明治の初期くらいまでのあいだに作成された文書を，古文書と呼ぶことにする．古代から中世までに作成されて現在に伝わっている文書数は，約25万通といわれているが，これに近世を加えると古文書は無数にあるといってよい．

これらの古文書の多くは，各地の文書館などに収集され保管されているが，その量があまりに膨大なため，どのような古文書をどれだけ所有しているのかを把握すらできていない文書館もある．ましてや，それらのすべてを翻刻し，あるいは電子化して，歴史研究の史料として利用できる形にするまでには，膨大な労力と時間が必要なのが現状である．

古文書に書かれた文字の特徴は，第1にその多くは毛筆で書かれていること，第2につづけ字が多いこと，第3にくずし字が多いことことの3点に集約される．もちろん，古文書の様態は書かれた時代や種類によってさまざまであるから，すべての古文書がこれらの特徴を持っているわけではない．古文書のなかにも活字印刷に近いような，読みやすい木版印刷物もある．しかし，未翻刻のものが圧倒的に多い近世の文書に限っていえば，おおよそ上記のような特徴を持っているといってよいだろう．

第1の毛筆であるという特徴によって，文字を構成する線の「かすれ」や「つぶれ」，運筆による線の濃淡が生じる．とくに線が「かすれ」たり「つぶれ」たりすることは，文字認識の処理を施すうえで重大な問題になる．第2のつづけ字であるという特徴によって，これまでの日本語手書き文字認識の技術を応用するためには，つづけ字のなかから1文字を切り出す必要が生じる．これが第3のくずし字であるという特徴と重なって，文字切り出しだけをとっても容易に解決できない難問が，古文書の文字認識の前に立ちはだかっている（図1.1）．

しかしながら，このようにたいへん困難に思える古文書の文字認識にも，研究に着手するためのいくつかの手がかりがある．まず対象とする文書の年代についていえば，未翻刻の文書の多さを考えれば江戸時代の近世文書にターゲットを絞ってよいだろう．近世に書かれた文書にも，江戸幕府の公式記録から個人の日記まで，さまざまなものがある．われわれは，歴史研究での重要性を勘案して，公的な記録文書を対象にしている．この種の文書は，毛筆書きされたものがほとんどである．おそらく技術的な容易さからいえば，木版刷りの板本を対象にしたほう

図 1.1: 古文書の文字（かすれ，つぶれ，つづけ字，くずし字が同時に現れる例）

が良い成果を期待できるだろう．しかし，われわれはあえて困難な毛筆手書きの文書の文字認識に挑戦している．

近世の公的な記録は「御家流」と呼ばれるくずし字によって書かれてある．つまり，文字のくずしの作法にはある程度の統一性がある．さらに，文書の種類によっては定型文が頻出する．たとえば，借金証文の場合ならば「申候處実正也」（もうしそうろうところじっしょうなり）といった語句がよく使用され，本文の最後は必ず「依而如件」（よってくだんのごとし）で結ばれる．用紙のどのあたりにどのような情報が書かれているかのレイアウトも，文書の種類によってははっきりとした構造を持っている．

これらのことを手がかりに，古文書の文字認識という遠大な研究に対してどのように取り組んでいるのかを，以下にご紹介したい．

1.2 プロジェクトの経緯

歴史学研究においては，古文書の翻刻が研究プロセスの重要な基礎的作業である．古文書翻刻作業は高度に知的な作業で，歴史の基礎知識，文書の種類やレイアウトに関する知識，定型文言・慣用表現の知識，文字の異体字やくずし方に関する知識と翻刻経験の蓄積が必要であり，人間が古文書翻刻作業をひととおりこなせるようになるまでには，相当の訓練期間を必要とする．古文書翻刻の知的プロセスを解明し，その知見にもとづいて古文書翻刻作業の一部を支援するシステムがあれば，歴史学研究の有効なツールとして活用しうるかもしれない．

研究プロジェクトの発足当時を振り返ってみると，古文書の文字認識をにらんだ研究は，文献 [1, 2, 3] など，ごくわずかしか発表されていなかった．これらの先行研究はいずれも，古文書文字認識の可能性を検証したにすぎないもので，本質的な技術的課題について解答を示したものではない．古文書翻刻支援システム実現のための，基本的かつ特殊な技術的課題に以下のものがある．

1. 古文書文字認識の技術　－　古文書特有の毛筆くずし字，つづけ字の辞書と認識．
2. 文書形式・定型文言の認識技術　－　近世文書に特有の文書類型，「恐々謹言」「仍而如件」などの頻出熟語の考慮．
3. システムと人間のインタラクション技術　－　古文書文字認識において人間が与える情報の範囲，認識結果の提示法，誤り修正方法など．

これらは従来の日本語手書き文字認識研究では未開拓の内容で，あらたな技術開発が必要な分野である．

上記の個別技術課題に関しては，共同研究者のひとりである柴山が，科学研究費基盤研究「東洋学における大量マルチメディア情報の提供方式の研究」（平成 7～8 年）で基礎的検討をおこなった．そこでは，歴史史料を対象にした画像資料の入力とデータベース化，ネットワークによる文字テキストや画像資料の提供方式についての研究の一部として，(1) ビデオ撮影による古文書の効率的画像入力法とコンピュータ上での史料復元，(2) 古文書画像の文字切り出しと文字認識に関する基礎的検討をおこなった．

また科学研究費補助金特定領域研究「人文科学とコンピュータ」（平成 7～10 年）のイメージ処理計画研究，公募研究において，山田，原，小島，川口が，劣化した古文書の画像処理，古文書のひらがな・漢数字に関する文字

認識研究を実施し，文書を限定したひらがなにおいて 65.8 %，漢数字において 92 %の文字認識率を得ている．

以上のような個別的な古文書認識技術に関する研究成果をもとにして，平成 10 年 8 月 5〜6 日に国際日本文化研究センターにおいて「第 1 回古文書ＯＣＲ（自動読み取り）シンポジウム」が開催された [4]（資料編を参照のこと）．同シンポジウムでは共同研究者等が研究発表をおこない，日本史・古文書学研究者，手書き文字認識研究者ら約 60 名が参加し，(1) 歴史研究者からみた古文書ＯＣＲへの期待，(2) 古文書ＯＣＲ研究の現況，(3) 日本語手書き文字認識の最先端技術の 3 つのテーマについて討議をおこなった．このシンポジウムの結果，当面の研究方略として以下の 4 点推進することで，参加者の意見の一致をみた．

1. 対象の選択において，書体の安定した公文書であり歴史的価値のたかいものを対象にする．
2. 文字認識のための辞書構築を進めるために，標準文字データベースを作成する．
3. 古文書読解に関する専門知識を整理し，システム化する．
4. 人間と機械の作業分担を明確化し，両者を円滑につなぐ知的ユーザインタフェースを構築する．

日本語手書き文字認識の最新技術を展開的に応用しつつ，上記課題の (1)〜(3) を達成し、課題 (4) であげられた知的ユーザインタフェースを備えた，古文書翻刻支援システムの開発をめざした研究の必要性が認識されている．

1.3 目的と概要

1.3.1 プロジェクトの目的

本プロジェクトの目的は，古文書翻刻支援システム開発に向けて，文字データベースなどの必要な研究環境の整備とシステム実現のための基礎的な検討を実施することにある．システム実現のための技術的なアプローチは，つぎの 3 点にある．

1. 古文書学の専門家が持つ古文書認識における認識過程をモデル化し，古文書解読のメカニズムを実証的にあきらかにする．
2. 日本語手書き文字認識技術を古文書に対して展開的に応用する．
3. 古文書翻刻支援に真に有効なマン・マシンインタフェースを検討する．

専門家の古文書解読プロセスをモデル化することは，知能情報学研究として興味深いテーマであるばかりでなく，その知見を利用することにより，古文書解読訓練方法の開発や支援ツールの開発にもつながる．古文書文字認識は，すでに性能向上の限界点に達している日本語手書き文字認識技術研究に，あらたな展開を与えうるものでもある．人文科学研究の現場で使用するコンピュータという観点からは，人間とコンピュータの作業分担のありかたを具現化する部分として，インタフェース研究が重要である．

本プロジェクトは，文字のくずしのはなはだしい文書を含むすべての古文書の解読や，古文書解読の完全自動化を目指すものではない．古文書解読プロセスのモデル化とシステムへの実装を通して，古文書解読という高度な知識処理過程を実証的に解明することと，同一文型・書体の文書が大量にあるような古文書の翻刻において，人間の作業負荷軽減に有効なシステム，人間が得意とする作業は人間が，機械が得意とする作業は機械がおこない，両者の円滑なインタラクションが確保できるシステムの開発が狙いである．

1.3.2 プロジェクトの概要

本プロジェクトの眼目は，つぎの 3 点にあるといえる．

1. 古文書専門家がもつ古文書解読の専門知識を構造化し，モデル化する．

2. 30年来の研究の蓄積を有する文字認識技術，なかでも日本語手書き文字認識に関する近年の飛躍的研究成果をもとに，文字認識の範囲を近世（江戸時代）古文書にまで展開して適用する．
3. 文字認識機能と古文書読解の専門的知識を内蔵した知的インタフェースを構築し，翻刻作業に関する習熟度のひくい作業者であっても，短時間によりおおくの翻刻作業がおこなえるシステムを開発する．

そのために当面必要となる作業には，つぎのようなものがある．

○ 古文書解読のための専門的知識の抽出と構造化．
○ 文字認識に必要な古文書文字認識用辞書の作成．
○ 古文書文字認識のアルゴリズム検討．
○ これらの作業を実施するための，基本的ツール群の開発．

具体的には，行書体および一部草書体を含み，語彙や文言が限られた証文・触書を中心とした近世文書を対象に，古文書文字認識のための辞書の作成，近世文書のレイアウト・頻出慣用表現などに関する専門的知識の構造化，古文書文字認識エンジンの開発，知的インタフェースの開発をおこなう．当面対象とする近世文書は，以下の文書である．

○「伏見屋善兵衛文書」（以下「伏見屋文書」と略す）（大阪市立大学所蔵）
○ 陸奥国会津郡小松川村「宗門改人別家別書上帳」（以下「宗門改帳」（しゅうもんあらためちょう）と略す）（個人蔵，年齢表記部分）

本プロジェクトの意義は，以下の点にある．

1. 古文書解読のための専門知識をモデル化することで，人間の知能情報処理を解明できる．
2. 日本語手書き文字認識の手法を，古文書に拡大適用するための方法論を確立できる．
3. 知識処理と文字認識を統合した，知的インタフェースのプロトタイプが作成できる．

1.4 古文書文字データベース

古文書文字認識の研究を進めるためには，研究者間で共有可能な研究の土台となる文字データベースが必要である．ところが，過去には古文書文字に関してそのようなデータベースは存在しなかったため，われわれはまずデータベース整備から作業をはじめた．古文書文字認識の試験データとなる文字データベースは，以下の観点から作成した．

1. 用例データとともに文字データが提供でき，知識処理を加えた文字認識の開発に供せられるもの．
2. 歴史研究上の汎用性のたかい文書からの文字．
3. 字種が限られているが，さまざまな筆跡のサンプルが多数得られるもの．
4. 標準的な古文書文字辞典の文字．

1の観点からは，大阪市立大学所蔵の「伏見屋善兵衛文書」を取り上げ，そこに登場する全文字の切り出しとデータベース化をおこなった．2の観点からは，『柳営日次記』の利用について検討を進めてきたが，データベース作成の費用と時間が莫大になるため，作業の重点を他の3つに絞ることとした．3の観点からは，「宗門改帳」に記載されたいくつかの文字のデータベース化をおこなった．4の観点からは，古文書翻刻者が利用する標準的な辞書のひとつである，東京堂出版『毛筆版くずし字解読辞典』を選択し，収録されている文字のデータベース化を完了した．

これまでに作成・公開した古文書文字データベースは，表1.1の通りである．これらのデータベースはすべて，HCRプロジェクトのホームページからダウンロードすることができる．

表 1.1: 古文書文字データベース HCD シリーズ

名称	内容	採字元	字種	文字数	画像
HCD1	年齢表記文字	宗門改帳	16	3,066	2 値
HCD1a	単位表記文字	宗門改帳	16	3,200	2 値
HCD1b	単位表記文字	宗門改帳	8	1,600	2 値
HCD1c	親族関係表記文字	宗門改帳	8	1,600	2 値
HCD1d	村役人表記文字	宗門改帳	8	1,456	2 値
HCD1e	貸地に関する文字	宗門改帳	8	1,600	2 値
HCD2	借金証文標題行	伏見屋文書	200 行	1,378	2 値
HCD2a	借金証文標題行	伏見屋文書	200 行	1,378	256 階調
HCD2b	借金証文標題行	伏見屋文書	200 行	1,378	24bit カラー
HCD3	借金証文標題文字	伏見屋文書	183	4,933	2 値
HCD4	借金証文全文文字	伏見屋文書	1,436	142,663	2 値

1.4.1 「伏見屋善兵衛文書」全文文字データベース

知識処理と組み合わせた古文書翻刻支援を考えた場合，定型文言が頻出するタイプの文書に焦点をあてることが有効である．近世の金子借用証文などは，文書の様式や文言が定型であり，当初の研究対象とするには最適であると判断した．われわれは，上記の条件を満たし種々の権利上の問題もクリアできる研究対象文書として，大阪市立大学が所蔵する「伏見屋善兵衛文書」（以降「伏見屋文書」）（図 1.2）を選択した．

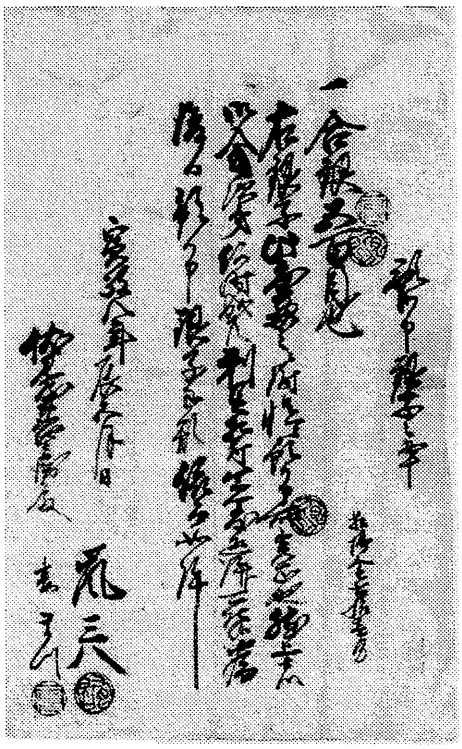


図 1.2: 「伏見屋善兵衛文書」

「伏見屋文書」は，大阪の元伏見坂町（現在の大阪市南区坂町）の茶屋，伏見屋善兵衛家に伝わった文書である．

伏見屋善兵衛は，遊興の地である伏見坂町のなかでも最大の茶屋として栄えた．また町年寄をつとめ，芝居興業にも関係し，何軒かの貸家をもち，金融業を営んだ．本文書は，文化から慶応年間にいたる各種の証文類である．芝居関係では，天保年間を中心に歌舞伎役者の芝翫，我童らの手附証文がある．伏見屋の金融・借家，同家内部の親族関係に関する諸証文・議定等も含まれている．文書の総数は，証文類が約 1,300 である．

文書からの文字切り出しとデータベース化は，つぎのような手順で実施した．

1. 文字認識実験の正解情報を作成するために全文を翻刻
2. 原文書をスキャナーでディジタル化し紙にプリント
3. プリントされた文書に対し，手作業でカラーマーカーを使って文字ひとつひとつを丸で囲む
4. マーク済みシートをスキャナーで再デジタル化
5. 文字切り出しプログラムで文字を切り出す
6. 翻刻データと照合しながら校正
7. マークの位置座標を利用してマーキング前画像から文字を再切り出し
8. 同一字種を集めて不良文字を削除し，翻刻文字と再照合

手順 2 でマークされたシートは，図 1.3 のようなものになる．われわれは，このシートから丸で囲まれた領域を自動的に切り出すプログラムを開発した．標題部分について文字を切り出し，文字データと照合した結果を図 1.4 に示した．

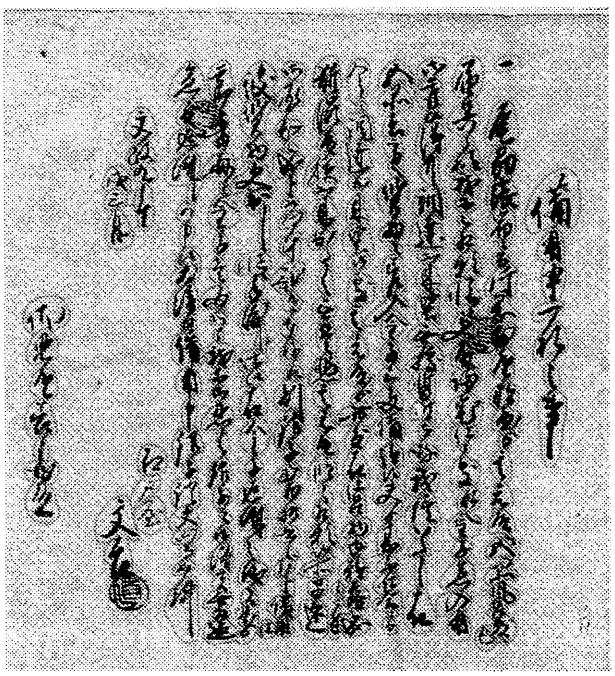


図 1.3: 文字部分をマークしたシート

「伏見屋文書」の全標題 4,995 文字から作成した古文書文字認識用データベースは，HCD3 の名称で公開している（表 1.2）．

また，「伏見屋文書」の全文をもとにした，約 14 万文字のデータベースも HCD4 の名称で公開の準備をしている．

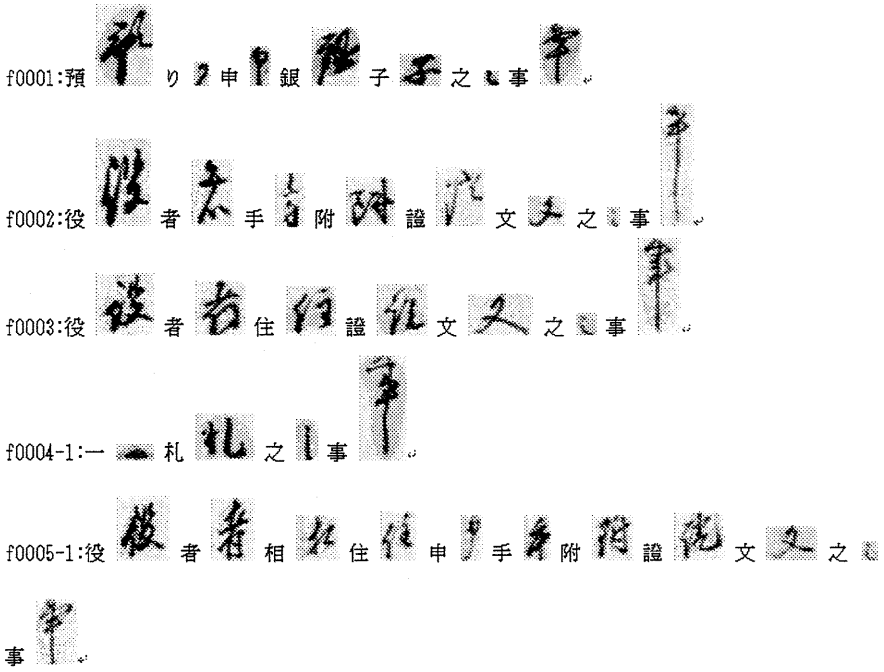


図 1.4: 文字切り出し結果

表 1.2: 「伏見屋文書」標題の頻出文字

字種	出現頻度	累積%
之	653	6.8
事	645	13.4
申	348	17.0
り	307	20.2
子	306	23.4
預	290	26.4
金	274	29.2
覚	270	32.0
文	256	34.6
證	229	37.0
一	225	39.3
請	211	41.5
屋	183	43.4
札	182	45.3
銀	156	46.9
月	154	48.5
年	136	49.9
家	125	51.2
状	124	52.5
借	105	53.6
通	103	54.6

1.4.2 「宗門改帳」文字データベース

われわれは，字種が限られているがさまざまな筆跡のサンプルが多数得られる文字データベースとして，共同研究者の川口洋が収集した「宗門改帳」記載文字のデータベース化を実施している．現在これらのデータを HCD1 (Historical Character Database 1) という名称で公開し，古文書文字認識の基礎実験に供している．HCD1 のシリーズに収録されている字種とサンプル数は，表 1.3〜1.8 のとおりである．

表 1.3: HCD1 収録の字種とサンプル数

字種	サンプル数	字種	サンプル数
ツ	200	八	200
一	200	九	200
二	200	十	200
三	200	壱	200
四	200	弐	200
五	200	年	200
六	200	拾	200
七	200	廿	66

表 1.4: HCD1a 収録の字種とサンプル数

字種	サンプル数	字種	サンプル数
田	200	両	200
畑	200	分	200
高	200	朱	200
石	200	家	200
斗	200	軒	200
升	200	間	200
合	200	馬	200
金	200	疋	200

表 1.5: HCD1b 収録の字種とサンプル数

字種	サンプル数
内	200
男	200
女	200
人	200
〆	200
長	200
横	200
夕	200

表 1.6: HCD1c 収録の字種とサンプル数

字種	サンプル数
父	200
母	200
子	200
悴	200
祖	200
弟	200
娘	200
房	200

表 1.7: HCD1d 収録の字種とサンプル数

字種	サンプル数
村	200
名	128
主	128
組	200
頭	200
百	200
姓	200
代	200

表 1.8: HCD1e 収録の字種とサンプル数

字種	サンプル数
借	200
貸	200
質	200
地	200
方	200
より	200
同	200
断	200

1.4.3 くずし字辞典文字データベース

「伏見屋文書」や「宗門改帳」といった実際の古文書から採字してデータベース化することも重要であるが，古文書文字辞典に登場するような典型的なくずし字のパターンをデータベース化することも有用であろう．われわれは多くの古文書翻刻者が利用している標準的な辞書のひとつである，東京堂出版『毛筆版くずし字解読辞典』[5] を選択し，出版社の許諾を得てそのデータベース化を実施した．

データベース化した文字は，同辞典のなかの「付録」を除く本編と増補のかな文字部分全 308 頁に登場する文字と用例，25,202 文字（用例も 1 文字とした）である．すべての文字および用例について，画像ファイル名，S-JIS コード，今昔文字鏡コード，読み，今昔文字鏡文字画像への URL を文字データとして作成し，くずし字画像を 400dpi の 2 値で画像取り込みした．

同時に，『毛筆版くずし字解読辞典』掲載文字の筆順情報を，タブレットＰＣを使って入力するツールを開発し，筆順の点列データを作成する作業を進めている（図 1.5，1.6)。

われわれはさらに，古文書翻刻者にとって必須の辞書になっている『くずし字用例辞典』[6] の電子化の作業も進めている。

残念ながら，著作権上の理由により当データベースを公開することはできないが，これを活用して後述の古文書文字認識研究，電子化古文書文字辞典の研究を進めている．

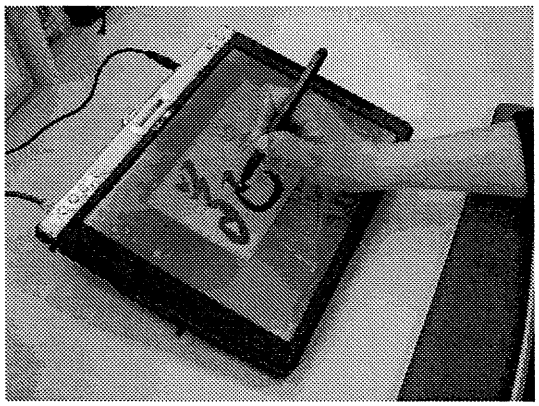


図 1.5: 筆順入力ツール

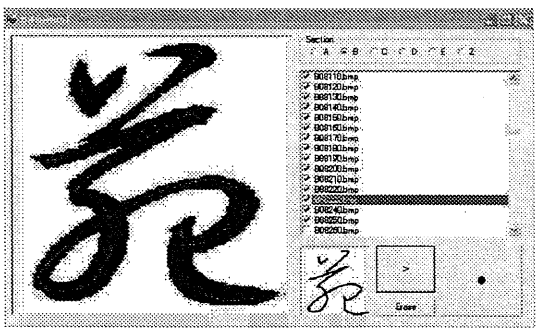


図 1.6: 筆順入力画面

1.4.4 文字切り出し研究用データベース

古文書のつづけ字のなかから 1 文字を切り出すことができたならば，手書き文字認識の技術を適用しやすくなる．ところがつづけ字から正確に文字を切り出すことは，至難である．文字切り出し自体が HCR のおおきな研究テーマでもある．文字切り出し研究を進めるためには文字の場合と同様，標準的なデータベースを整備して多くの研究者がおなじ土俵で議論ができる環境を整える必要がある．

われわれは，文字切り出し研究用データベースとするために，「伏見屋文書」から標題行を抽出した．ノイズが比較的すくなく 1 行のみからなる標題で，複数の文字から構成され，かつ文字がつづけ字になっている 200 標題を選択して，そのフルカラー画像および翻刻文字をデータベース化し，HCD2 の名称で公開をしている（図 1.7）．

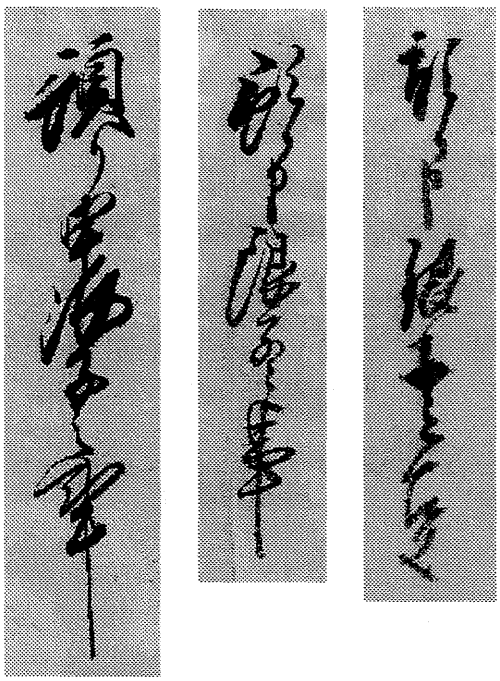


図 1.7: 文字切り出し研究用データベース収録画像の例

1.5 古文書用例データベース

古文書に登場する文面の用例を収集することによって，そこから知識を抽出し，その知識を使った古文書翻刻支援が可能となる．またその用例は，定型的な文言が頻出するタイプの文書を収集するのが効果的である．古文書文字データベース作成の対象とした「伏見屋文書」は，そのほとんどが金子借用証文である．証文類は「実正也」「急度返済可申候」「依而如件」などの定型文言が多く見られ，文書の様式も安定しているため，用例データベースの対象として最適である．われわれは，古文書文字データベース作成作業と平行して「伏見屋文書」全文約 243,000 文字を翻刻し，用例データベースとした．作成された用例データベースは，後述の「知識による翻刻支援」研究に利用している．

1.6 古文書文字切り出し

古文書文字の切り出し，及び文字認識の基礎的研究をおこなうために，古文書標題のみを対象とした文字パターン辞書データベース構築と，関連するユーザインターフェイスの開発を実施した [7]．古文書の形態は縦横の長さ，おおきさが一様でないため，古文書レイアウトの把握や他の古文書との比較が容易にできない．そのため古文書概略画像をピラミッド型の上位層で抽出し，その抽出した抽象化レベルのレイアウトから標題部分だけに着目して原画像から標題部分の抽出をおこなった．

古文書画像のピラミッド型によるレイアウト抽出をおこない，その結果を判断し，標題の抽出を射影ヒストグラム法とラベリング法のふたつの手法を用いておこなった．その結果，78 %の割合で標題抽出をおこなえ，形式が未知である文書の分類が会話型で短時間におこなえるユーサインターフェースを開発した（図 1.8,1.9）．しかし，印影や裏写りの影響を受けたものに対しては，本手法では解決されず，また誤って文字の一部分のみ抽出されたものもある．文字の一部分のみ抽出された文書に対する改善は，今後各閾値を一定値から各画像の画素値の分布に対して変化させた実験をおこないたいと考えている．また，古文書画像において，レイアウトを認識するルール，及びその実現する手法について考察した．今後このレイアウト認識の実験もおこないたいと考えている．

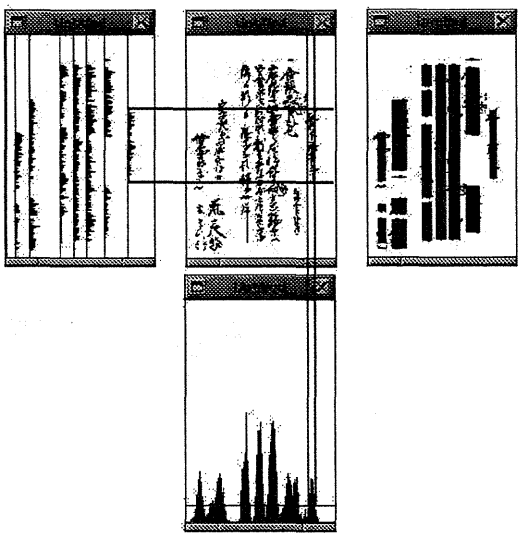


図 1.8: ヒストグラムによる抽出範囲選択

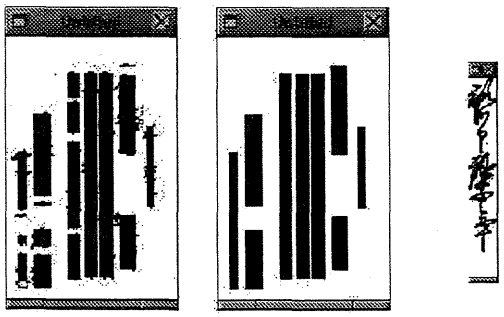


図 1.9: 文字列の抽出個所及び標題抽出結果

1.7 古文書文字認識

従来の文字認識過程には，つぎのような特徴がある．

1. 切出しから認識までが順次処理される
2. 辞書への正規化では失われる情報がある
3. 文字サイズ・意味カテゴリーなどをパラメタにした辞書検索をおこなっていない
4. 通常は、認識過程の終了後の後処理で整合性がチェックされる

こうした従来型の認識プロセスにおいて，人間の文字認識プロセスに近いモデル化が可能かどうかを検討した [8]．具体的には，

1. 各文字パターンのサイズなどの特徴が失われない方法
2. 辞書検索時にサイズ等のパラメタが指定できる
3. 後処理から認識へバックトラックする機能
4. 文字切出しと認識の同時処理がおこなわれる方法

などを検討する必要がある．

以下に示す文字認識の実験では，上記の 1,4 について実現した．正規化は，認識しようとする対象画像に対して，文字パターン辞書から取り出されたパターンを対象画像のサイズに一致するように変換することである．われわれは，従来の認識プロセスとはまったく逆の発想で検討した．

まず，2-gram を用いた切出し，及び認識プロセスについて検討した．

1. 標題の先頭文字に出現する文字カテゴリーに含まれる 1 文字パターンを辞書から取り出す．
2. つぎに対象画像の文字幅を、辞書から取り出した文字パターン幅に変換する．すなわち正規化する．
3. つぎにマッチングに移行する．マッチングは重ね合わせ法によるが，隣接文字の「侵入」や「連結」を切出すためにマッチングをおこなう範囲を限定しなければならない．このために，マスク処理をおこなう．
4. 対象画像上での探索範囲は，おおむね経験則から文字パターンの高さの 2 倍としている．
5. マッチングにより，両パターンの距離が一定のしきい値以下になったとき，一致したとみなす．
6. 一致したパターンで対象画像のパターンを消去し，これがつぎの対象画像となる．

以上があらたな試みの認識プロセスの概要である．この実験結果から，2-gram を用いて切出し・認識をおこなった場合，約 90 %の認識率を得た（図 1.10,1.11）．この方式は，従来の人間の動作に比較してより近いのではないかと考えている。

このほかにもわれわれは，非線形正規化によりすくない文字サンプルから多様な文字サンプルを生成する手法や，手書き文字入力からくずし字辞典を検索するウェブインタフェース（図 1.12）についても研究を進めている．また HCD1 を対象とした自己想起型ニューラルネットを使った古文書文字認識で，未知パターンに対する平均認識率 99.06 %を達成している [9]．

1.8 知識による翻刻支援

翻刻時に遭遇する読めない文字（不明文字）の前後文字から n-gram の情報を使って不明文字の正解候補を提示する可能性について検討した [10]．用例データとして「伏見屋文書」を使用し，翻刻支援手法の検討と検証をおこなった．その結果，前後の既知文字から 3-gram および 2-gram の情報を使って不明文字の正解を検索する実験により，第 10 候補までで 72.70 %の正解率を得られると推定できた．

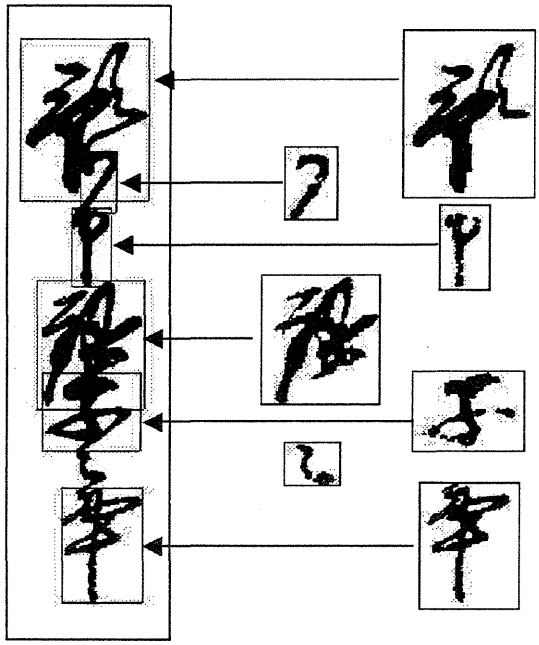
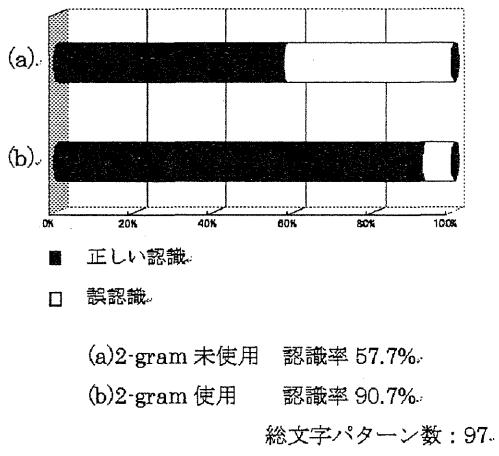


図 1.10: 切り出し・認識結果の例



(a)2-gram 未使用　認識率 57.7%

(b)2-gram 使用　　認識率 90.7%

総文字パターン数：97

図 1.11: 切り出し・認識結果

本手法を Microsoft Word のマクロとして実装し，GetAMoji マクロの名称で公開している（図 1.13）．翻刻文を Word に呼び出し，GetAMoji を実行すると「□」文字の部分の正解候補が提示される．GetAMoji の利用試験をおこなったところ，翻刻経験のない初心者が辞書なしで翻刻した結果の正解文字数が有意に増加することがわかり，システムの有効性が確かめられた．

GetAMoji には「伏見屋文書」から作成した近世借金証文用辞書がサンプル辞書として付いているが，利用者が翻刻文の Word ファイルから，自分の辞書を作成する機能も持っている．

本手法は，不明文字の前後の文字が正しいと仮定して，その情報から不明文字の候補を提示するものである．したがって，前後の文字がそもそも誤っていたり，文字数の推定が誤っていたり，不明文字が連続してしまった場合には，正しい候補文字の提示ができない．本手法の応用として，英文のスペルチェックに対応するような，翻刻済み文字に対する検証システムのようなものも考えられるだろう．また本手法は，証文類という一定の表現が頻出するパターンをとる文字列に対して有効な手法であって，その他の種類の文書対してこの手法がどの程度有効で

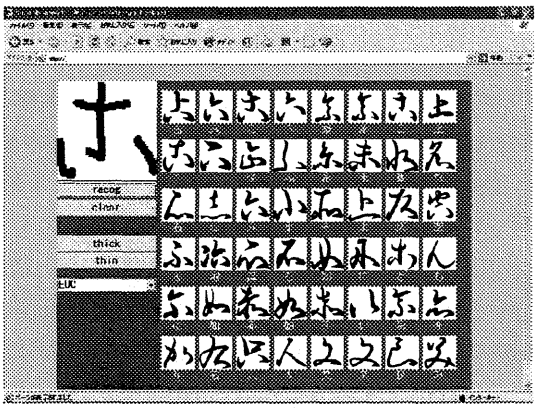


図 1.12: 手書き文字入力からくずし字辞典を検索するウェブインタフェース

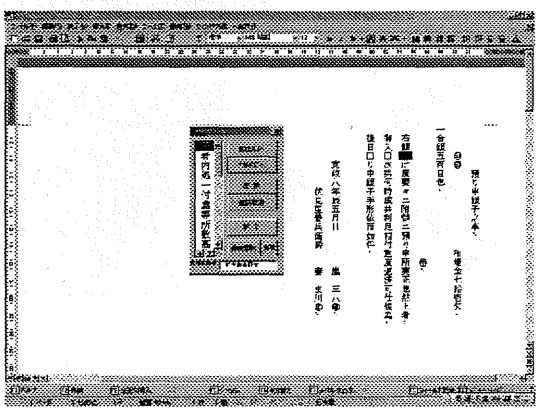


図 1.13: GetAMoji マクロ

あるかは今後の検討が必要である.

なお，GetAMoji の Web 版も作成し，HCR プロジェクトのホームページから公開している（図 1.14）.

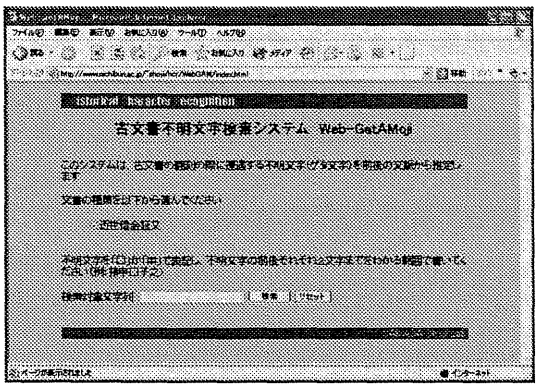


図 1.14: Web 版 GetAMoji

1.9 電子化古文書文字辞典

翻刻者が古文書を翻刻する際には，古文書文字辞典を参照しながら作業を進める．古文書翻刻作業に使われている標準的な辞典のひとつである『毛筆版くずし字解読辞典』 [5] は，文字の第 1 ストロークの方向から検索でき

るという，ほかの辞典にみられない特長を有している．しかしながら紙ベースの辞典では，その検索の利便性はかならずしもたかいとはいえない．

われわれは古文書文字データベース作成作業において同辞典をディジタル化している．そこで同時点のディジタル情報を使って，紙の辞典よりも検索性をたかめた電子化古文書文字辞典の開発を進めている．電子化古文書文字辞典では，従来の「漢字」や「読み」からの文字検索に加えて，文字の外形や運筆からの検索を可能にする．

われわれは，ある文字と第 1 ストローク方向が同一で，しかも外形が似ている類似文字を検索する機能をもった，Windows 環境で動く電子化古文書文字辞典を開発した（表 1.15）．さらにわれわれは，オンライン文字認識技術を応用して，運筆から検索できる電子化古文書文字辞典の開発にも取り組んでいる．

将来的には，電子手帳のような携帯型のツールに電子化古文書文字辞典を搭載することを目指している．

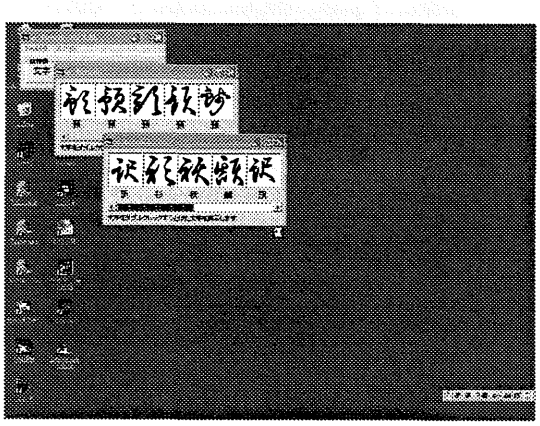


図 1.15: 電子化古文書文字辞典 eKuzushi

1.10 おわりに

平成 11 年度より開始した「古文書翻刻支援システム開発（ＨＣＲ）プロジェクト」のこれまでの成果の概要は，以上のとおりである．現在までのところ，古文書文字データベース，古文書用例データベース，および知識による翻刻支援システムについて研究成果を公開するにまで到っている．古文書文字切り出し，古文書文字認識，電子化古文書文字辞典についてもデータを整備と平行して基礎的研究と試験システムの開発を進めている．

HCR プロジェクトのホームページは，

http://www.nichibun.ac.jp/~shoji/hcr/

である．最新の研究成果報告や本報告で述べた成果物の公開は，当ホームページからおこなっている．

第 2 章

古文書文字データベース

2.1 HCD1 シリーズ

HCD1(Historical Character Data 1) シリーズは，帝塚山大学経営情報学部の川口洋氏によって作成された古文書文字データベースを収録したものである．HCD1 には古文書の一種である宗門改帳（しゅうもんあらためちょう）から採字した年齢表記文字，単位文字，親族関係表記文字が収録されている．文字画像データは，後述のように 2 値の PBM アスキー形式の画像ファイルを連結した形で提供している．データの提供サイトは，つぎのとおりである．

http://www.nichibun.ac.jp/~shoji/hcr/

データの使用条件等については，川口氏のホームページ http://kawaguchi.tezukayama-u.ac.jp/を参照されたい．

2.1.1 歴史研究上の意義

徳川幕府によって，享保 6（1721）年から 6 年に 1 度づつ実施されていた「子午改め」と呼ばれる調査によれば，日本の総人口は 18 世紀を通じて停滞していたが，19 世紀中期からゆるやかに増加を始めた．ことに北関東，東北地方では，1 世紀におよんでいた人口減少が，19 世紀前期を底として増加に転じた．このような持続的人口成長の開始は，伝統社会から近代社会への移行を端的に示す指標の一つと解釈される．

現在のところ，持続的人口成長がどのような地域社会の状況下で始まり，明治以降に継続していくのか，という極めて素朴な課題については，試論の域を出ていない．他方，江戸時代における民衆の生活は，家族構造，出産力などの基礎的な側面で，地域差に富んでいたことが近年改めて指摘された．したがって，近代移行期における民衆生活の理解を深めるには，時系列的変化に加えて，個別集落の地域的特色を全国的展望のなかに位置づけ，地域差の生じた要因を解明する歴史地理学の研究方法が有効と思われる．

このような課題を追求するには，各地に保存されている古文書史料を組織的に収集，蓄積，分析する研究方法を開発することが求められる．

2.1.2 史料の概要

記載内容

江戸時代の日本では，「宗門改帳（しゅうもんあらためちょう）」と総称される古文書史料が，17 世紀末から 19 世紀中期の明治初年まで，全国で作られていた．たとえば，陸奥国会津郡，大沼郡，下野国塩谷郡（現在の福島県南会津郡，大沼郡，栃木県塩谷郡）の一部を含む南山御蔵入領（みなみやまおくらいりりょう）では，元禄 7（1694）年あるいは元禄 8 年から明治 3（1870）年まで毎年，村ごとに名主の手によって作成され，代官所と自宅に 1 部づつ保管されていた．

南山御蔵入領に所属する小松川村（福島県南会津郡下郷町）には，散逸した 9 年分を除いて，寛政 4（1792）年から慶応 4（1868）年に至る 77 年間の「宗門改人別家別書上帳」が保存されている．この史料には，以下に示す画像のように，記載単位ごとに，旦那寺の本末関係，所在地，宗派，旦那寺の名称，持高，質地，家屋規模，屋根の材料，構成員の名前，筆頭者との続柄，年齢，異動，牛馬数，世帯規模などが記録されている．史料的性格を検討すると，南山御蔵入領の「宗門改人別家別書上帳」は，現住人口を世帯単位に記録した史料であり，婚姻，養子縁組，奉公などの異動が生じてから史料に登録されるまでの期間は，多くの場合 1 年以内であったことが確認できる．

「宗門改帳」が，継年的に保存されている村では，史料的制約に十分留意すれば，人口変動のほかにも，初婚年齢，死亡年齢，養子や婚姻による人口移動の範囲といった人口再生産構造に影響を持つ人口学的指標を長期間にわたって求めることができる．さらに，家族形態，家族周期，相続や改名に関する慣習など，民衆生活の具体像を示す情報を知ることも可能である．

保存状況

国外の研究者のなかには，このように豊富な内容を持つ「宗門改帳」の存在に注目して，日本を宝島 (treasure island) と呼ぶ者もいるようである．「宗門改帳」の全国的な所在調査は，現在進行中であるため，ここでは思い切って，南山御蔵入領の保存状況を全国に普遍化することにより，人口学的分析に耐える史料が保存されている村の数を推計したい．

南山御蔵入領には 271 ヶ村が所属している．このうち，50 年以上にわたって毎年の「宗門改人別家別書上帳」が保存されている村は，陸奥国会津郡石伏村，鴇巣村，金井沢村，小松川村，大窪村，寺山村，寺村，沢入村，大沼郡桑原村の 9 ヶ村である．したがって，271 ヶ村のうち約 3 %の村で長期間の人口学的指標を求めることができる．

全国で 50 年以上にわたる「宗門改帳」が保存されている村の割合を，ひとまず南山御蔵入領と同様 3 %と仮定する．『天保郷帳』には，北海道を除いて天保 5（1834）年の日本には，63562 の村が記録されている．したがって，全国で 1900 余りの村において，長期間にわたる「宗門改帳」が保存されていると推計される．天保 5 年における平均的な村の人口規模は，約 420 人なので，記録されているのは延べ 3990 万人，和紙 1 枚に約 10 人分が記録できるとすると，史料の枚数は合計 399 万枚と見積もることができる．

「宗門改め」の制度は藩によって多様であり，史料の保存状況も地域差が大きい．6 年に 1 度しか「宗門改め」を実施しなかった水戸藩，紀州藩などの諸藩もみられる．これに対して，南山御蔵入領は，日本でも有数の史料の宝庫である．そのため，実際に「宗門改帳」が保存されている村は，1900 ヶ村を相当下回ると思われる．それにしても，近代的国家の成立以前に，50 年以上の期間にわたって数百ヶ村もの人口現象を分析できる国は，おそらく日本に限定されると思われる．質，量ともに，宝島と呼ばれるのにふさわしい史料を作成，保存してきたのが日本

社会の特色のひとつである．

2.1.3 文字データベース化

画像データの入力

「宗門改帳」をはじめ和紙に筆墨で書かれた古文書の画像情報をデジタル化する方法として，①イメージ・スキャナーを用いて取り込む，②デジタル・カメラで撮影する，③写真撮影したうえで，フィルム・スキャナーを用いて取り込む，という3種類があげられる．検討の結果，鮮明な画像データを比較的廉価で作成できる点，史料の保存機関に持ち込む機器が簡便である点，過去に撮影されたフィルムをデータベース構築の資源として継承することができる点などを考慮して，③の方法を用いて古文書の画像情報をデジタル化するのが妥当と判断した．

具体的には，カメラ（NIKON F2，NIKOR 55mm/F3.5）を照明台の上に固定して，ハロゲンランプで照明を当て，FUJICOLOR SUPER G ACE 400 のフィルムで，史料の見開き2ページを1画像として写真撮影した．次に，フィルムを PHOTO-CD に書き込んだ．PAINT SHOP を用いて，PHOTO-CD から 1536 x 1024 DOTS の解像度で画像を読み込み，グレイスケール（256 階調）に調整，ノーマルフィルター（シャープ強）をかけ，1世帯を1画像に編集した後，JPEG 形式で保存した．小松川村の 75 年におよぶ「宗門改人別家別書上帳」は，1692 画像，411.9MB，1画像平均 243.4KB の容量で保存された．

実験対象文字

漢数字で表記される年齢，牛馬数，世帯規模，持高，家屋規模といった情報のなかで年齢は，結婚年齢，出産年齢，死亡年齢，夫妻の年齢差，年齢別人口構成，生命表といった人口学的指標を算出する場合，とくに重要な基礎的情報となる．年齢を表記した漢数字の種類は限定されるうえに，古文書史料には，世帯構成員の名前の下のほぼ固定した位置に記録されているため，セグメンテーションも比較的容易と予測される．

「文政八年酉年二月　宗旨家別人別分限書上帳　小松川村，寺山村，大久保村，沢入村，寺村」のうち小松川分の史料には，273 種類，3898 文字が使われている．このうち，たとえば HCD1 に採録した年齢を表記した 16 種類の古文書文字（ツ，一，二，三，四，五，六，七，八，九，十，壱，弐，年，拾，廿）は，全体の約 22 %に相当する 868 文字出現する．とくに，「弐」，「壱」，「四」は，出現順位が 10 位以内に入る頻出文字である．加齢などにともない史料作成年次ごとの文字の出現頻度は変化するが，16 種類の文字は常に頻出する．

採字

「宗門改帳」古文書画像データベースに登録されている古文書画像から，実験対象となる文字に外接する枠をかけて手作業で切り出し，2値化してビットマップ画像として保存する，という手順で採字した．実験対象文字のうち「廿」を除いた文字を各 200 個づつ採字した．「廿」については 66 個しか採字できなかった．

77 年間にわたる小松川村の「宗門家別人別改書上帳」のうち，寛政 4（1792）～寛政 12（1800）年の名主は多蔵，享和 2（1801）～文政 6（1823）年の名主は太郎兵衛，文政 7（1824）～安政 4（1857）年の名主は忠左衛門，安政 5（1858）～慶応 4（1868）年の名主は忠右衛門である．史料の作成責任者は，この 4 人であるが，書き役などが書類を書く場合もあるため，実際の執筆者は特定できない．採字した文字には，複数の人物が書いた文字が含まれていることだけは確実とみられる．

古文書文字は，和紙に毛筆で書かれた手書き文字である．一種類の文字であっても，字形，字体に相当なばらつきがみられる，続け字（連綿体）が多用されている，文字の太さが多様である，前後の文字などの影響で，文字の

大きさが多様であるといった特徴を持っている．そのため，古文書読解技能を持つ研究者であっても誤読を犯す場合がある．

2.1.4 HCD1 技術情報

収録字種とサンプル数

ツ	200
一	200
二	200
三	200
四	200
五	200
六	200
七	200
八	200
九	200
十	200
壱	200
弐	200
年	200
拾	200
廿	66

データベースフォーマット

Record1	Record2	Record3	
PBM File	PBM File	PBM File	

サンプル文字の画像ファイルが PBM フォーマット（アスキーエンコーディング）で 3,066 文字分ならんでいる．

レコードフォーマット

Line1	Magic Number	P1
Line2	Comment	# CharacterID SJIS JIS(ASCII)
Line3	Size	Width Height
Line4-	Binary Image	ASCII Encoded Binary Image

Line1 は P1 で固定．P1 は PBM フォーマットでは 2 値画像のアスキーエンコーディングを意味する．Line2 はコメント行であるが，ここに文字 ID，SJIS コード，アスキーエンコーディングされた JIS コードがブランクで区切られて入っている．Line3 は画像サイズで，画像の幅と高さがブランクで区切られて入っている．Line4 以後に文字の 2 値イメージがアスキーコードで入っている．0 は白画素，1 は黒画素である．

文字 ID のネーミングルール

ツ	0-nnn
一	1-nnn
二	2-nnn
三	3-nnn
四	4-nnn
五	5-nnn
六	6-nnn
七	7-nnn
八	8-nnn
九	9-nnn
十	A-nnn
壱	B-nnn
弐	C-nnn
年	D-nnn
拾	E-nnn
廿	F-nnn

nnn は文字内で通し番号

2.1.5 HCD1a 技術情報

収録字種とサンプル数

田	200
畑	200
高	200
石	200
斗	200
升	200
合	200
金	200
両	200
分	200
朱	200
家	200
軒	200
間	200
馬	200
疋	200

データベースフォーマット

HCD1 に準じる.

レコードフォーマット

HCD1 に準じる.

文字 ID のネーミングルール

田	10-nnn
畑	11-nnn
高	12-nnn
石	13-nnn
斗	14-nnn
升	15-nnn
合	16-nnn
金	17-nnn
両	18-nnn
分	19-nnn
朱	1A-nnn
家	1B-nnn
軒	1C-nnn
間	1D-nnn
馬	1E-nnn
疋	1F-nnn

nnn は文字内で通し番号.

2.1.6 HCD1b 技術情報

収録字種とサンプル数

内	200
人	200
男	200
女	200
〆	200
長	200
横	200
夕	200

データベースフォーマット

HCD1 に準じる.

レコードフォーマット

HCD1 に準じる.

文字 ID のネーミングルール

内	20-nnn
人	21-nnn
男	22-nnn
女	23-nnn
𠂉	24-nnn
長	25-nnn
横	26-nnn
夕	27-nnn

nnn は文字内で通し番号.

2.1.7 HCD1c 技術情報

収録字種とサンプル数

父	200
母	200
子	200
悴	200
祖	200
弟	200
娘	200
房	200

データベースフォーマット

HCD1 に準じる.

レコードフォーマット

HCD1 に準じる.

文字 ID のネーミングルール

父	28-nnn
母	29-nnn
子	2a-nnn
悴	2b-nnn
祖	2c-nnn
弟	2d-nnn
娘	2e-nnn
房	2f-nnn

nnn は文字内で通し番号.

2.1.8 HCD1d 技術情報

収録字種とサンプル数

村	200
名	128
主	128
組	200
頭	200
百	200
姓	200
代	200

データベースフォーマット

HCD1 に準じる.

レコードフォーマット

HCD1 に準じる.

文字 ID のネーミングルール

村	30-nnn
名	31-nnn
主	32-nnn
組	33-nnn
頭	34-nnn
百	35-nnn
姓	36-nnn
代	37-nnn

nnn は文字内で通し番号.

2.1.9 HCD1e 技術情報

収録字種とサンプル数

借	200
貸	200
質	200
地	200
方	200
より	200
同	200
断	200

データベースフォーマット

HCD1 に準じる.

レコードフォーマット

HCD1 に準じる.

文字 ID のネーミングルール

借	38-nnn
貸	39-nnn
質	3a-nnn
地	3b-nnn
方	3c-nnn
より	3d-nnn
同	3e-nnn
断	3f-nnn

nnn は文字内で通し番号.

2.2 HCD2 シリーズ

HCD2 は，古文書文字切り出し研究のために作成されたデータベースである．大阪市立大学所蔵『伏見屋善兵衛文書』から比較的ノイズの少ない 200 標題（1,378 文字）を選択し，その文字画像と標題の翻刻文字情報を収録している．

収録されているのは文化年代から慶応年間にいたる各種の証文類の標題で，芝居関係の手附証文，金融・借家・伏見屋の親族に関する諸証文・議定等が含まれている．

HCD2 には，翻刻文字情報の csv ファイルと文字画像ファイルが含まれている．データベース名と内容の対応は，つぎのとおりである．

名称	内容	画像ファイル形式
HCD2	2 値画像	PBM
HCD2a	階調画像	PGM
HCD2b	フルカラー画像	JPG

画像ファイル名	翻刻文
001	預り申銀子之事
002	預り申銀子之事
003	預り申銀子之事
004	家質證文之事
005	預り申銀子之事
006	預り申銀子之事
007	預り申銀子之事
008	預り申銀子之事
009	預り申銀子之事
010	家質利銀請負證文之事
011	預り申金子之事
012	預り申金子之事
013	借用申金子之事
014	預り金證書之事
015	預り申金子之事
016	引當借用金證文之事
017	譲り證文之事
018	預り申銀子之事
019	預り申金子之事
020	引当借用證文之事
021	証金光寺
022	差入申證文之事
023	上嶋屋善右衛門
024	年賦證文之事
025	預り申金子之事
026	預り申金子之事
027	譲り證文之事
028	預り申銀子之事
029	預り申銀子之事
030	預り申銀子之事
031	預り申金子之事
032	預り申銀子之事
033	預り申銀子之事
034	預り申銀子之事
035	年賦證文之事
036	預り申金子之事
037	借用申金子之事
038	預り申金子之事
039	預り申金子之事
040	印鑑
041	家質證文之事
042	家屋敷質流し證文之事
043	預り申銀子之事
044	質物請状之事
045	質物請状之事
046	質物請状之事
047	質物請状之事
048	質物請状之事
049	乍恐口上
050	両替取引通請負一札之事

画像ファイル名	翻刻文
051	一札之事
052	家屋鋪永代賣渡證文之事
053	元伏見坂町居宅家證文
054	家附物譲り渡一札
055	永代賣渡申家屋鋪之事
056	付物代請取一札之事
057	家屋敷帳切賣券證文之事
058	家附物賣渡一札之事
059	家附物賣渡一札之事
060	家附物賣渡一札之事
061	譲り渡證文之事
062	譲り渡一札之事
063	預り申銀子之事
064	親類請一札之事
065	借屋請状之事
066	借家請状之事
067	家附物借り受一札之事
068	座敷借り受證文之事
069	座敷借請負一札之事
070	親類請負一札之事
071	親類請負一札之事
072	親類請負一札之事
073	親類請負一札之事
074	親類請負一札之事
075	親類請負一札之事
076	親類請負一札之事
077	親類請負一札之事
078	親類請負一札之事
079	親類請負一札之事
080	座敷借請負一札之事
081	親類請負一札之事
082	親類請負一札之事
083	親類請負一札之事
084	親類請負一札之事
085	親類請負一札之事
086	親類請負一札之事
087	借家請状之事
088	借家請状之事
089	借家請状之事
090	借家請状之事
091	借家請状之事
092	貸家請状之事
093	差入申一札
094	親族受一札
095	親類請負一札之事
096	親類請負一札之事
097	親類請負一札之事
098	親類請負一札之事
099	親類請負一札之事
100	親類請負一札之事

画像ファイル名	翻刻文
101	親類請負一札之事
102	座敷借り受負一札之事
103	座敷借受負一札之事
104	親類請負一札之事
105	座敷借受負一札之事
106	借家請状之事
107	親類請負一札之事
108	親類請負一札之事
109	借家請状之事
110	借家請状之事
111	借家請状之事
112	借家請状之事
113	借家請状之事
114	貸家請状之事
115	貸家請状之事
116	貸家請状之事
117	貸家請状之事
118	借家請状之事
119	借家請状之事
120	家附物譲り渡證文之事
121	借家請状之事
122	借家請状之事
123	親類請負一札之事
124	借家請状之事
125	家賃銀諸事請状之事
126	約定一札
127	借家請状之事
128	家附物譲り渡一札
129	借家請状之事
130	親類請一札之事
131	借家請状之事
132	借家請状之事
133	借家請状之事
134	借家請状之事
135	親類請負一札之事
136	借家請状之事
137	村物書付申事
138	借家請状之事
139	借家請状之事
140	借家請状之事
141	借家請状之事
142	借家請状之事
143	借家請状之事
144	借家請状之事
145	借家請状之事
146	借家請状之事
147	借家請状之事
148	宗旨手形之事
149	宗旨寺請状之事
150	宗旨手形之事

画像ファイル名	翻刻文
151	宗旨手形之事
152	宗旨手形之事
153	宗旨手形之事
154	宗旨手形之事
155	宗旨手形之事
156	宗旨手形之事
157	宗旨寺請状之事
158	宗旨手形之事
159	宗旨寺請状之事
160	宗旨手形之事
161	宗旨手形之事
162	宗旨手形之事
163	宗門手形之事
164	宗旨手形之事
165	宗旨手形之事
166	宗旨手形之事
167	宗旨手形之事
168	宗旨手形之事
169	差入申一札之事
170	請取一札之事
171	議定一札之事
172	片身分ケ一札之事
173	為取替議定一札之事
174	為取替議定一札之事
175	差入申一札之事
176	譲渡證文之事
177	預り申銀子年賦證文之事
178	譲り證文之事
179	差入申一札之事
180	差入申一札之事
181	差入申頼一札之事
182	家督相續頼状之事
183	預り申年賦銀之事
184	預申銀子之事
185	御影御請待志
186	宗旨手形事
187	宗旨手形之事
188	宗旨手形之事
189	一札之事
190	永代経料一札之事
191	乍憚口上
192	年賦證文之事
193	親類請負一札之事
194	親類請負一札之事
195	親類請負一札之事
196	親類請負一札之事
197	酒造元建
198	口上
199	御割附
200	印鑑

2.3 HCD3 シリーズ

HCD3 は、古文書文字認識研究のために作成されたデータベースである。大阪市立大学所蔵『伏見屋善兵衛文書』について、不明文字がない 900 標題の全 4,933 文字（184 字種）を 1 文字づつ切り出して、その 2 値画像と翻刻文字情報を収録している。

HCD3 の文字画像の 1 部は、HCD2 に収録された標題行と元データレベルでおなじものであるが、HCD2 と HCD3 とでは画像加工工程が異なるので、両者はピクセルレベルで完全に一致するものではない。

HCD3 には，翻刻文字情報の csv ファイルと PBM 形式の文字画像ファイルが含まれている.

文字画像ファイル名は、以下のような命名規則になっている。

9999-999-999-99#99.pbm

#以前の 15 桁の数字がおなじ文字は同一の標題行からの文字、#以下の 2 桁で先頭文字からの番号をあらわしている。たとえば、

0001-000-000-00#01, 預
0001-000-000-00#02, り
0001-000-000-00#03, 申
0001-000-000-00#04, 銀
0001-000-000-00#05, 子
0001-000-000-00#06, 之
0001-000-000-00#07, 事

ならば、文書番号 0001-000-000-00 の標題は「預り申銀子之事」になる。

文字切り出し作業の都合上、元画像ファイルをいったん紙にプリントしたものを再スキャンした。そのため、文字の輪郭は原文書の品質を保っていない。印影やシミ、つづけ字の連結部分については手作業で除去した。文字のかすれ、虫食いはそのままノイズとして残してある。また 2 値化の後にメディアンフィルタをかけて、輪郭線の荒れを平滑化した。

第 3 章

古文書画像の標題文字切り出し

3.1 はじめに

計算機技術の進歩に伴い，人文学分野においても工学的手法が取り入れられ，研究が進められている．そのひとつとして古文書画像のデータベース化が挙げられる．古文書画像データベースの検索においては，標題，発信人，受取人，年代などの目録を作成し，その目録より対象とする画像を検索するのが一般的である．さらに全文検索をおこなうには，翻刻，解題，読み下し文のテキストが必要となる．しかしながら目録作成等をすべて手作業で行うには膨大な時間と費用，専門的知識を必要とする．古文書文字の切り出し，認識の研究は，それらの作業を軽減するのに大いに貢献するに違いない．

本研究は，古文書文字の切り出し，および文字認識の基礎的研究をおこなうために，古文書標題のみを対象とした文字パターン辞書のデータベース構築と関連するユーザインターフェイスの開発を目的にしている．古文書画像は「伏見屋文書」の約 1,300 文書，2,000 画像を対象にする．

3.2 古文書画像の抽象化

古文書の原画像をピラミッド構造により，抽象化して概略画像を得る．ピラミッド構造とは，原画像に対してピラミッドの上位層で画像を抽出する方法である．概略画像を抽出する理由は，

1. 縦または横に長い古文書画像のレイアウトの把握
2. 文字列の位置関係，様式，形態の把握
3. レイアウト特徴による文書の分類

が容易にできるためである．

3.3 射影ヒストグラム法による標題抽出

3.3.1 ヒストグラム

つぎに概略画像からの行，及び文字列の抽出の概要を図 3.1 に示す．抽出された概略画像 $(m \times n)$ より垂直射影ヒストグラム ν_i をとる．ν_i は，

$$\nu_i = \sum_{j=0}^{n-1} p(i,j) \qquad (i = 0, 1, \ldots, m-1) \tag{3.1}$$

で表される．ここで i は概略画像の水平 (x) 方向位置，j は同画像の垂直 (y) 方向とする．

行抽出のために式 (3.1) に基づく閾値を $t_c(t_c \geq 0)$ とし，$\nu_i \geq t_c$ の条件を満たす連続した変数 i を縦書き 1 行と定め，$i_s \leq i \leq i_e$ の範囲を抽出する．ここで，i_s は連続したの始点，i_e は終点を表す．ここで，抽出された各々の行を k（ただし，$k = 1, 2, \ldots, k_{max}$）とする．

つぎに得られたそれぞれの行 k に対して水平射影ヒストグラム h_{kj} をとる．h_{kj} は，

$$h_{kj} = \sum_{i_s}^{i_e} p(i,j) \qquad (j = 0, 1, \ldots, n-1) \tag{3.2}$$

で表される．この h_{kj} が得られた個所を文字部分として定め，連続したヒストグラムの上端を行の始端，下端を行の終端とする．

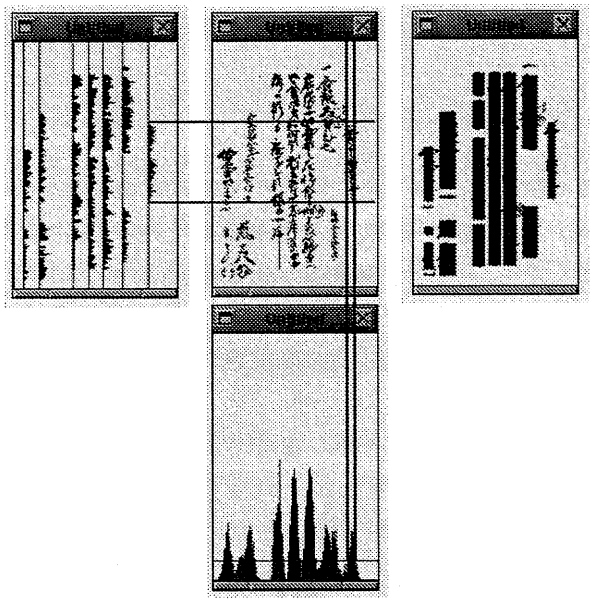


図 3.1: ヒストグラムによる抽出範囲選択

3.3.2 標題抽出

ヒストグラムによって抽出個所を決定したが，図 3.2（a）に示すように本来，標題や差出人等の意味のある文字列の一部分で空白が出来ているため，このままでは文字列として抽出できない．そのために必要に応じて補間操作をすることとした．この結果を図 3.2（b）に示す．

つぎに補間した抽出範囲から標題を抽出する際のルールは，①文書の最右端の行を標題と仮定する．標題の抽出方法は，②最右端の行より抽出個所の矩形 4 隅の座標を概略画像上で取得する．③その座標を概略画像から原画像用に座標変換をおこない，④原画像の標題部分のみを読み取り，抽出する．図 3.3 に抽出結果を示す．

3.3.3 実験結果

全画像 1987 枚に対して標題が抽出できたのは 712 枚，全体に対して抽出できた割合は 36 %である．しかし抽出できなかった画像数の中には封筒，裏書などもともと標題が存在しない画像が 993 枚含まれている．それらを全体から除き，標題が存在している画像だけで考えると抽出できなかった画像は 282 枚である．よって標題が存在する画像だけで考えると，標題が存在する画像 994 枚に対して，標題が抽出できたのは 712 枚であり，72 %の割合という結果が得られた．

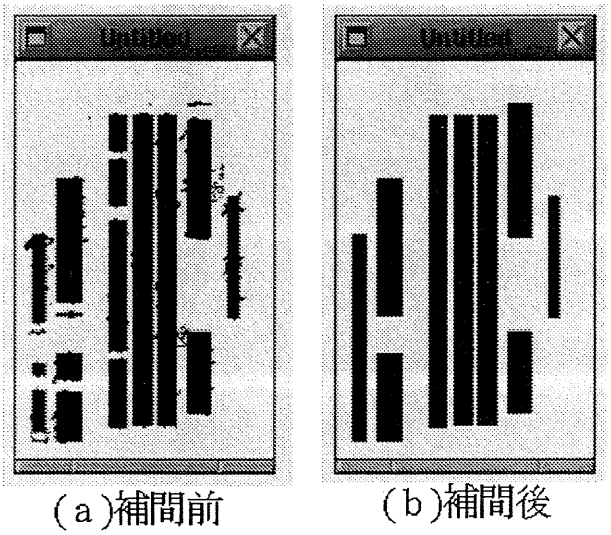


図 3.2: 文字列の抽出個所

図 3.3: 標題抽出結果

3.3.4 射影ヒストグラム法による行抽出の問題点

射影ヒストグラム法による行抽出の問題点は，第 1 に文字の一部が削れることである．垂直射影ヒストグラムでの閾値により，文字の一部分が欠ける．標題部分（文字列）としては認識できるが，抽出した文字列に対して文字認識をおこなう場合，文字の削れで正しい認識ができない．

第 2 に，文字列が傾いている場合，文字列の始端及び終端部分の垂直射影ヒストグラムの値が低くなり，文字列の始端及び終端の文字の一部が削れる場合がある．また，行間が狭い場合には，垂直射影ヒストグラム上において文字列の終端と隣の文字列の始端部分が重なってしまい，行間で分割する事が困難である．

第 3 に，図 3.4 に示すように隣の文字列からの侵入（図 3.4 左側：「申」の左側の印影，図 3.4 右側：「舞」の左側の印影）の影響がある．垂直射影ヒストグラムによって文字列と定めた範囲に，隣の文字列の文字の一部が侵入している場合，その侵入している文字の一部も抽出される．

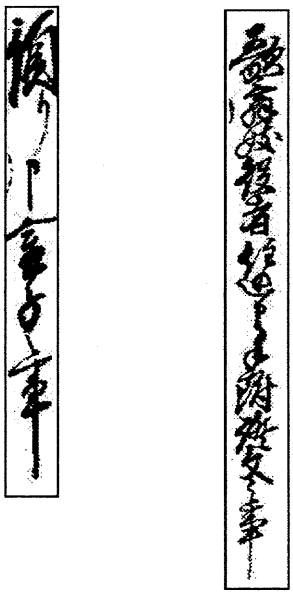


図 3.4: 隣接文字の侵入例

3.4 射影ヒストグラム法とラベリング法による標題抽出

3.4.1 ラベリング法による標題抽出

つぎに射影ヒストグラムの問題点を改善するためにラベリング法の併用を考える．概略画像よりラベリング法を用いて標題を抽出する．ラベリング法の利用は黒色，つまり文字部分をひとつの塊としてみることができるため，前章示した射影ヒストグラム法による行抽出の問題点が解決できる．

前処理

概略画像に対してそのままラベリングを行うと偏と旁，文字と文字がそれぞれ離れたいた場合文字にかすれがある場合に，ひとつの文字，行として抽出することが難しい．この手法は柴山 [11] がおこなった実験でも示されている．したがって，偏や旁，文字と文字など抽出した意味のある文字列をひとつの塊として把握するために，垂直射影ヒストグラムによる一定の閾値以上の範囲を目安として塗りつぶしによって文字間の接続をおこなう処理（以下，結合処理という）をほどこす．

ラベリング法

前処理をした画像に対してラベリング処理をおこなう．ラベリング法とは連結している全ての画素に対しておなじラベル（番号）を付け，異なった連結成分には異なったラベルを付ける処理である．ラベル付けをおこなうと同時に各ラベル（連結成分）のラベル枠

$$q_n = (i_{min}, j_{min}, i_{max}, j_{max}) \tag{3.3}$$
$$n : \text{ラベル番号 } (n = 1, 2, \ldots, m)$$

も求める．

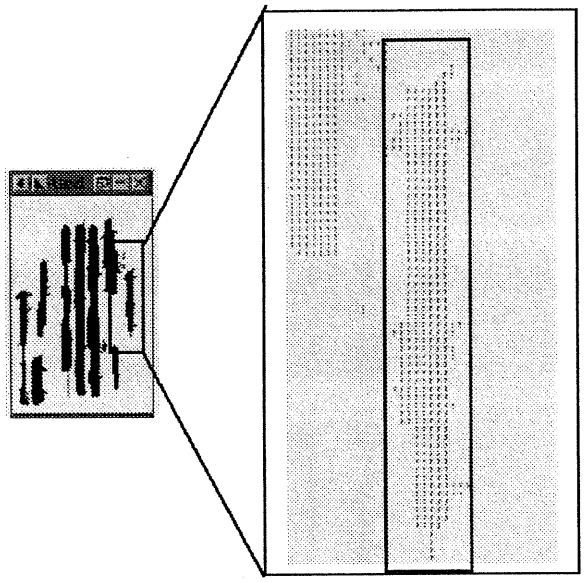


図 3.5: ラベリング処理画像とラベル枠

実験結果

1) 標題抽出例

図 3.6 (a) (b) は前節で示した隣接文字の侵入に関する問題に対して，隣接文字の侵入を抽出することなく標題のみを抽出できたことを示すものである．

2) 標題抽出不可例

図 3.7 は結合処理の際，垂直ヒストグラムの閾値が固定値によるために，標題文字が閾値以下になり結合処理が実行されなかったため，文字の一部しか抽出できていない．

3.4.2 ラベル枠を用いた抽出

ラベリング法のみでは前節図 3.7 のように標題の一部分のみ抽出されてしまう場合が生じる．そのためにラベリング法による抽出方法にラベル枠による抽出方法を合わせて標題の抽出を行う方法が考えられる．これはたとえば左右に分離した文字の一部分に外接する矩形を描き，その矩形が一部重なるような場合，同一ラベルを与えひとつの文字と見直す手法である [12], [13].

文字セグメントルール

ラベル番号が $n1, n2$ となる連結成分が存在するとき，そのラベル枠をそれぞれ前節の式 (2.3) より

$$\begin{aligned} q_{n1} &= (i_{n1min}, j_{n1min}, i_{n1max}, j_{n1max}) \\ q_{n2} &= (i_{n2min}, j_{n2min}, i_{n2max}, j_{n2max}) \end{aligned} \tag{3.4}$$

とする．ただし $n1 \leq n2$ とする．

このとき，q_{n1} に対して q_{n2} が以下の 3 つの条件を満たすとき，$n2$ のラベルを $n1$ に変換する．

$$j_{n1min} \leq j_{n2min} \leq j_{n1max} \tag{3.5}$$

図 3.6: 標題の抽出例

かつ

$$i_{n1min} \leq i_{n2max} \leq i_{n1max} \tag{3.6}$$

かつ

$$i_{n2max} - i_{n1min} \geq (i_{n2max} - i_{n2min})/2 \tag{3.7}$$

上記の各々は，図 3.8 (a) (b) (c) に対応する．または

$$j_{n1min} \leq j_{n2min} \geq j_{n1max} \tag{3.8}$$

かつ

$$i_{n1min} \leq i_{n2min} \geq i_{n1max} \tag{3.9}$$

かつ

$$i_{n1max} - i_{n2min} \geq (i_{n2max} - i_{n2min})/2 \tag{3.10}$$

ここで，式 (3.7) の場合，$n1$ の左端と $n2$ の右端の距離（図 3.8(c) 参照）を A，$n2$ の左端と右端の距離を B とする．条件 $A \geq (B/2)$ を満たすとき，つまり x 方向に対して $n2$ の領域が 1/2 以上 $n1$ に含まれるとき，ラベル変換をおこなう．

以上の手法を全ラベルに対して実施する．

図 3.7: 標題抽出不可例

実験結果

標題が存在する画像 994 枚のうち，前章で述べた射影ヒストグラムによる標題抽出によって標が抽出できなかった 282 枚を対象に，ラベリング法による標題抽出をおこなった．その結果 282 枚のうち 64 枚に関して標題を抽出する事ができた．

1) 標題抽出例

前節で示した古文書に対して，ラベリング法による標題抽出では標題の一部だけ抽出されていたのに対し，ラベル枠との併用による標題抽出では，正しく抽出できた（図 3.9）．

3.5 レイアウト認識

古文書画像において，標題，本文，日付，差出人，受取人等を認識するルール，及びその実現する手法について提案する．

3.5.1 行の定義

前章で求めたラベル枠を用いてそのラベル枠の左上の座標を (i_{min}, j_{min}) ，右下の座標を (i_{max}, j_{max}) とし，それによって求められる行 Q_n を

$$Q_n = (i_{min}, j_{min}, i_{max}, j_{max}) \tag{3.11}$$

n : ラベル番号

と定める．

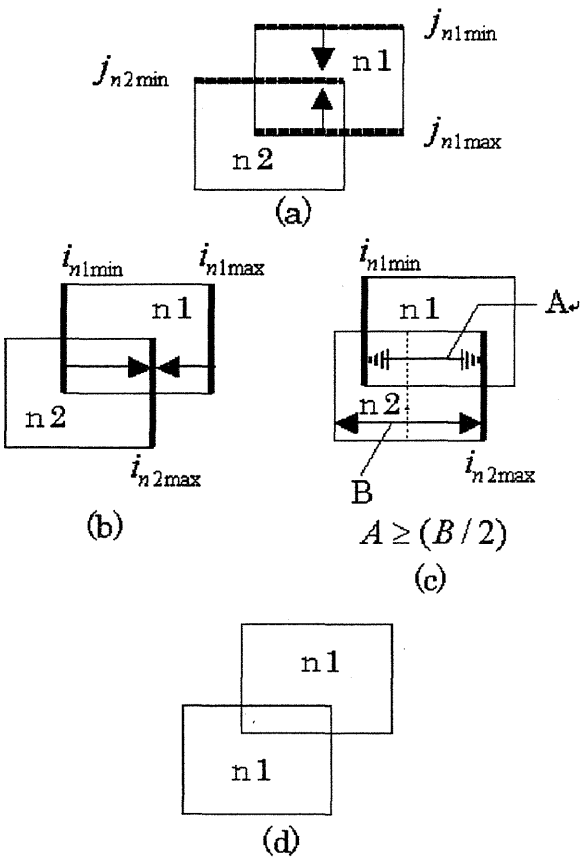


図 3.8: 文字切出しルール

3.5.2 認識ルール

各々のレイアウトを

注釈 1（標題より右側上部にある行）：C1
注釈 2（標題より右側下部にある行）：C2
標題：T,　本文：B,　日付：D,
差出人：S,　受取人：R,　追記：P

とする．これを要素という．

また，概略画像の水平射影ヒストグラムをとり，その上端と下端の中心を Y_2，上端と Y_2 の中心を Y_1，Y_2 と下端の中心を Y_3 とする（図 3.10）．ここで，$Y_1 = h/4$, $Y_2 = Y_1 + h/4$, $Y_3 = Y_2 + h/4$ である．ただし，水平射影ヒストグラムの上端と下端の距離を h，原点は左上隅とする．

以下のルールに基づきレイアウトを決定する（図 3.11）．

1) 注釈 1（C1），注釈 2（C2）

ラベル枠の座標とレイアウト分割ルールにおいて，

$j_{max} \leq Y_1$ のとき，$Q_n = C1$
$Y_3 \leq j_{min}$ のとき，$Q_n = C2$

とする．

図 3.9: 標題の抽出例

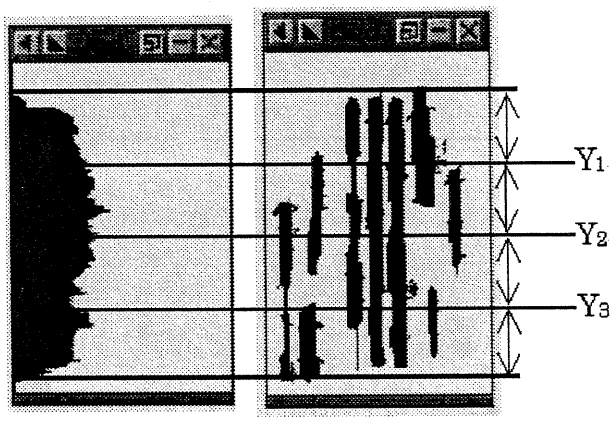


図 3.10: レイアウト認識基準

2) 標題 (T)

$Q_n = C1, Q_n = C2,$ の i_{min} をそれぞれ，$i_{C1min}, i_{C2min}, Q_n = C1, Q_n = C2$ を除く他の行 Q_n の i_{min} を i_{Omin} とすると，

$$i_{min} \leq i_{C1min}, \text{または} i_{min} \leq i_{C2min}$$

かつ

$$i_{min} \geq i_{Omin}$$

のとき $Q_n = T$ とする．

3) 本文 (B)

$j_{min} \leq Y_1$ かつ $Y_3 \leq j_{max}$ のとき，$Q_n = B$ とする．

4) 日付 (D)

$j_{min} \leq Y_1$ かつ $Y_2 \leq j_{max} \leq Y_3$ のとき，$Q_n = D$ とする．

5) 差出人 (S)

$Y_2 \leq j_{min} \leq Y_3$ かつ $Y_3 \leq j_{max}$ のとき，$Q_n = S$ とする．

6) 受取人 (R)

$Y_1 \leq j_{min} \leq Y_2$ かつ $Y_2 \leq j_{max} \leq Y_3$ のとき，$Q_n = R$ とする．

7) 追記 (P)

$Q_n = D, Q_n = S, Q_n = R$ の i_{min} をそれぞれ $i_{Dmin}, i_{Smin}, i_{Rmin}$ とすると，

$$i_{min} \leq i_{Dmin} \text{または} i_{min} \leq i_{Smin} \text{または} i_{min} \leq i_{Rmin}$$

かつ

$$j_{min} \leq Y_1 \text{かつ} Y_3 \leq j_{max}$$

のとき $Q_n = P$ とする．

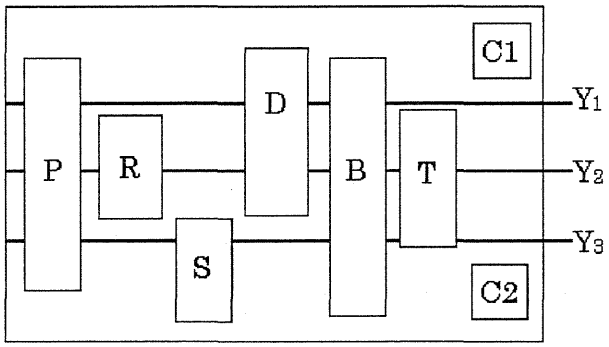


図 3.11: 認識基準と要素配置の関係

3.6 文字パターン辞書による文字セグメント方式

3.6.1 文字セグメント方法

抽出した標題画像より，文字パターン辞書を用いて文字セグメントをおこなう．文字パターン辞書とは，標題画像の各文字に対し，1 文字ずつに分割し，その文字がどの標題の文字で，どのような文字であるかをまとめたデータベースである．抽出した標題画像に対して，標題の先頭文字から文字パターン辞書の各文字を用いたテンプレートマッチングをし，マッチングが一致した場合，その文字を標題画像から切出し，そのつぎの文字に対しても同様にマッチングと文字セグメンテーションをする (図 3.12)．

3.6.2 実験方法

正規化

抽出した標題画像の各文字のおおきさと，文字パターン辞書の文字のおおきさは異なる．そのため，文字パターン辞書の文字を標題の文字のおおきさに合うように，正規化をする．

正規化の方法は，文字パターン辞書の文字のたかさの範囲内における，標題画像の水平方向の文字幅を検出し，そのなかでもっともおおきい幅を最大文字幅とする．つぎに文字パターン辞書に対しても同様に，水平方向の文

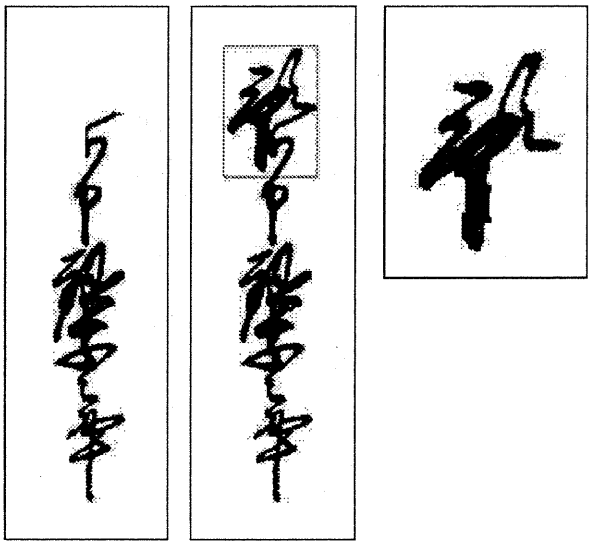


図 3.12: マッチング処理方法

字幅を検出し，さらに最大文字幅を検出する（図 3.13）．その標題画像側の最大文字幅と文字パターン辞書側の最大文字幅が合うように，文字パターン辞書を拡大縮小する．

n-gram による文字選択

文字パターン辞書から文字を選択する際の知識ベースとして，はじめに各標題の先頭文字にあたる文字から順にマッチングをおこなう．先頭文字の文字セグメントが終了し，つぎの文字に対してマッチングを行う際は，2-gram の情報を用いて，文字パターン辞書より順にマッチングする．

3.6.3 実験結果

今回は，抽出した標題画像のなかから，10 標題を選択し実験をおこなった．また文字パターン辞書に関しても，この 10 標題に対応したものを使用した．

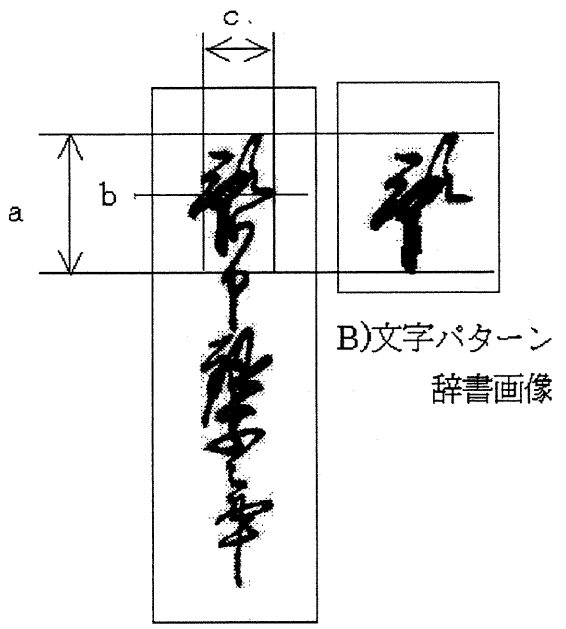
その結果，10 標題，97 文字のうち，82 文字マッチングに成功した．その成功した 82 文字のうちの 12 文字は他の標題から切出した同文字のパターン辞書によってマッチングが成功した．図 3.14 に実験結果を示す．

図 3.14 に関して，「之」の部分はマッチングできなかった．これは標題画像側の「之」がちいさく，また文字幅も狭いため，正規化が上手く出来なかったためである．また，今回の実験では「之」の文字に対してマッチングをする際，2-gram により「之」のつぎに来る語も文字選択の候補に入れマッチングをした．そのために，「之」でマッチングせずに，つぎに来る「事」でマッチングした．

その他に今回の実験での問題点として，標題画像からマッチングした個所を削除する際に，発生したノイズなどにより，その後の文字のマッチングに影響を及ぼしてしまう．また，異なった文字でマッチングをした場合，2-gram による文字選択が有効とならないという点をあげることができる．

3.7 おわりに

古文書画像に対してピラミッド型によるレイアウト抽出をし，その結果を判断して標題の抽出を射影ヒストグラム法とラベリング法の 2 つの手法を用いておこなった．その結果，78.1 %の割合で標題抽出を行え，形式が未



a：文字パターン辞書の文字の高さ
b：aの範囲内での最大文字幅位置
c：aの範囲内での最大文字幅

図 3.13: 文字パターン辞書の正規化

知である文書の分類が会話型で短時間におこなえるユーザインターフェースを開した.

また，古文書画像において，レイアウトを認識するルール，及びその実現する手法について考察した．現在，このレイアウト認識の実験を進めている.

文字パターン辞書による標題文字セグメントに関しては 10 標題に対して 84.5 %の割合でマッチングできた．今後他の標題画像にも同じ実験をおこない，また，切り取り，文字認識についても実験をつづける.

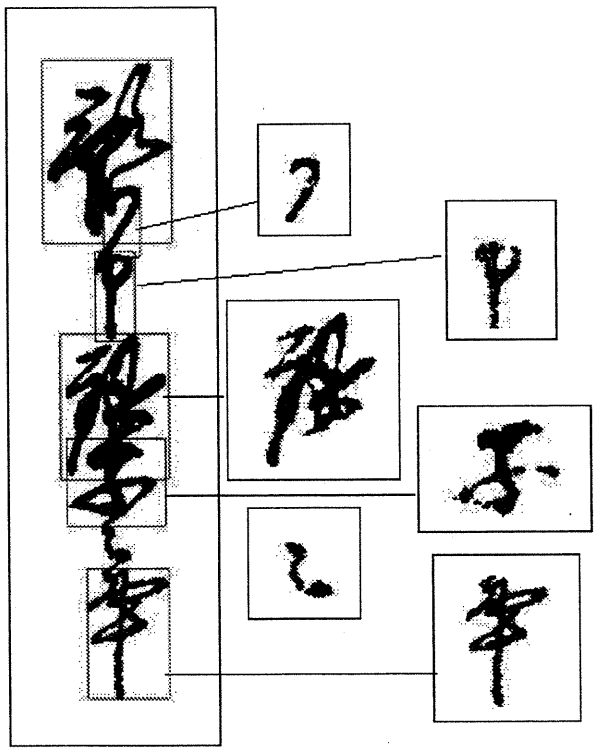


図 3.14: 実験結果と文字パターン辞書

第 4 章

古文書文字認識プロセスの検討

4.1 はじめに

毛筆によるくずし字やつづけ字で記述された古文書の解読は，個別のくずし字の判別や認識とともにことばや文章の判別・認識を行って，逐次に解読を進める．すなわち，文字，用語，文体の三位一体による解読が必要とされる．さらに文書の解読のためには，文書が記された時代背景や関連する知識が必要であることはいうまでもない．このうち，文字の識別に関する情報のみだけを取り上げても部首，扁，旁，画数，筆順，筆勢，筆圧と多くの要素が存在する．

一方，コンピュータによる古文書の文字認識をしようとする場合，入力画像から文字や語の画像特徴を抽出し文字認識へと進める必要があるが，そこでは形状，線，エッジ等による画像からの限られた特徴を基本に考えねばならない．これは，前述した人間がおこなう文書の認識プロセスとはおおきく異なる．さらに，古文書文字認識を従来型の文字認識モデルにあてはめると，文字切り出し，正規化などの過程で，前述の人間が解読の際に用いる情報のいくつかが失われると想定される．例えば，正規化によって文字の形状が変形し，極端な場合，本来の文字とは全く異なった他の字形との類似度が増すことになると推定される．これらの古文書画像から人間が得る特徴とコンピュータから得られる特徴の違いを図 4.1 に示す．

本報告では，従来型の文字認識モデルに従って古文書文字認識をおこなう場合の正規化が文字の類似性に与える影響について調べ，その結果について述べる．また，従来型の文字認識プロセスとは異なる古文書文字認識プロセスについて検討し，その実験結果について示す．

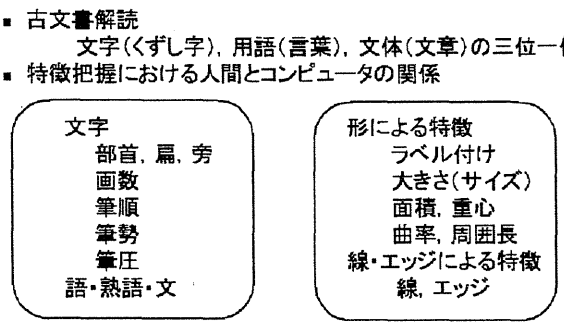


図 4.1: 古文書解読と文字画像処理

4.2 文字認識プロセスと古文書標題文字

4.2.1 文字認識過程と切り出し・正規化

手書き文字認識や漢字文字認識は，認識対象になる画像から2値化，レイアウト認識，ノイズ除去等の前処理がおこなわれ，個別文字が切り出される．この個別文字は，認識辞書とのマッチングのために正規化がおこなわれる．その後，文字の特徴抽出がされ，判別・認識がおこなわれる．このプロセスを，図4.2に示す．

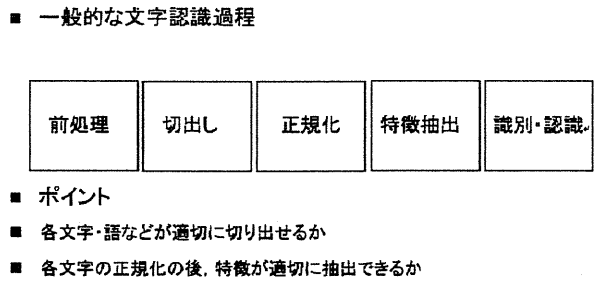


図4.2: 文字認識過程と切り出し・正規化

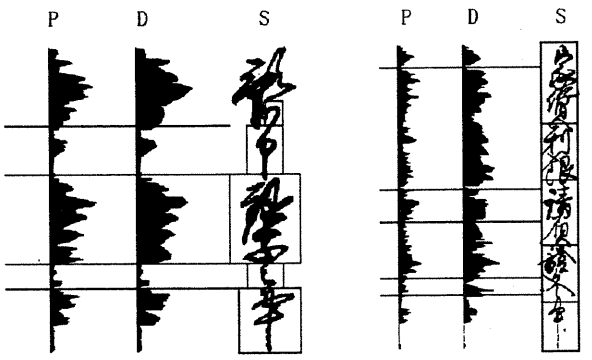
このプロセスで古文書を対象にする際には，各文字・語が適切に切り出せるか，また，正規化後の適切な特徴が抽出できるかが問題になる．

まず，切り出しについてである．従来から文字列の特徴を把握するのに水平方向の画素値に基づく射影ヒストグラム（図4.3：P）が用いられる [7] [14]．しかし，毛筆のつづけ字では，特徴を把握しにくい．そこで，最左端の画素から最右端の画素までの距離をヒストグラム化するとより特徴が掴みやすい（図4.3:D）．これから概ね，語を単位として切出せることが判る．関連する事例を図4.4，4.5に示す．なお，図4.3左側の『預り申銀子之事』の「預」と「り」，右側の『家質利限請負証文之事』の「文」と「之」が1字のように（侵入）重なっている．これらはおおきな問題で，HCRにおける最大の課題でもある．つぎに正規化の問題点について述べる．通常，文字認識に先だって正規化が必要になる．これは認識辞書の参照時にサイズ等を辞書の基準に合致させねばならない理由による．従って，正規化とは

- 位置　　文字の中心を移動
- 大きさ　　外接文字枠の幅・高さを伸縮
- 回転／傾き　文字主軸を所定の座標軸に
- 濃度　　平均濃度，最大／最小濃度
- 線幅　　線幅を所定の文字幅に変換

など所定の基準に移動・変換することである．サイズの正規化例を図4.6に示す．古文書文字認識においては，この正規化がまったく異なる字形との類似性に影響し，認識率を低下させる要因にならないかが問題となる．

つぎに，古文書文字の特徴を如何に把握するかである．漢字文字認識では，線素方向，周囲形状，パターン濃度分布等が用いられる．古文書文字では，図4.7に示す線素方向，標本点抽出が考えられるが，本報告では重ね合わせ法を用いて実験している．



P:画素射影ヒストグラム　D:距離射影ヒストグラム　S:文字列画像

図 4.3: 古文書標題文字列の特徴

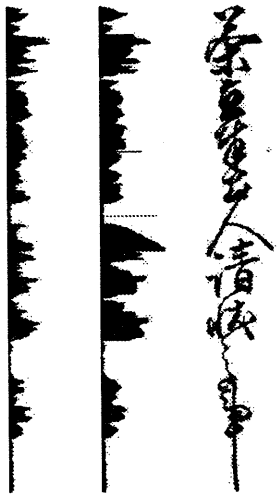


図 4.4: 標題事例 1

4.2.2 古文書文字パターン辞書の作成・構築

現在，われわれは古文書翻刻支援システムの開発のために，「伏見屋文書」に基づく文字パターン辞書の作成・構築を進めている．本報告に用いた文字パターンは辞書の構築過程で，本証文類の標題のみである．現在，本文の辞書構築も進捗中である．文字パターン辞書は，以下の手順で作成している．

まず，文書を複写した紙面（モノクロ画像）上で，人手により切り出す文字をピンクマーカペンで囲む．これをスキャナーで読み取り，画像処理を行って閉曲線で囲まれた部分（図 4.8:右上パターン [カラー画像]）を切出す．これを 2 値化して，1 文字のパターン（図 4.8:右下パターン [2 値画像]）にする．切り出した文字パターンを翻刻した文字とリンクした一覧が図 4.9 である．

図 4.10 は，図 4.9 に示す各文字パターンの属性情報である．ファイル名は図 4.9 に示す文字パターン画像で，字種は当該の翻刻文字である．図中の W,H は，文字パターンの各々幅，高さを表す．位置は，当該パターンの前後文字列の中での出現位置を示す．

文字パターン辞書として構築した標題のみの文字数は，文字種 193 種，4,622 パターンである．

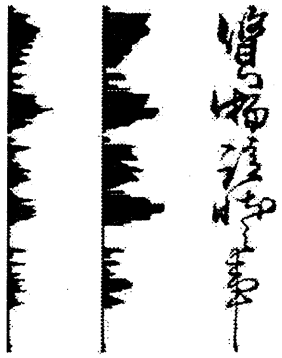


図 4.5: 標題事例 2

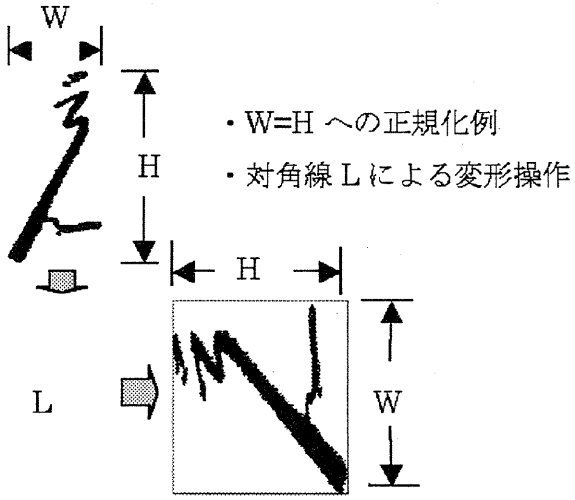


図 4.6: サイズ変換の正規化

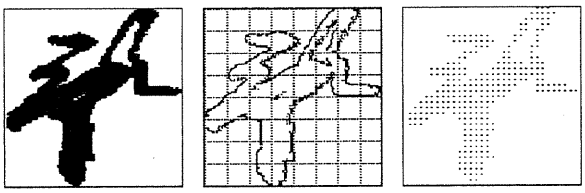
4.3 文字パターンの正規化と類似性

ここで文字認識過程における正規化が類似性にどの程度影響を与えるか．また，文字種と文字サイズに相関があるかなど正規化の問題点や文字の特徴について調べてみる．

4.3.1 文字パターンの特徴

図 4.11 では，文字パターンの W/H 比分布を表示している．W/H 比 r は，r=6.40〜0.098 である．

表 4.12 では，例えば「事」の文字パターンは，r が最大 1.78〜最小 0.098 であることを示し，出現頻度は 515，全体の文字数比は約 0.1 である．193 種の文字パターンから，「預」，「り」，「申」，「之」，「事」について，r による散布図を図 4.13 に，また重心による散布図を図 4.14 に示す．この結果，概ね W/H 比による字種の特徴が表れていることが判る．これは，重心による特徴よりバラツキがあるようにみえる．詳細は，さらに調べる必要があるが，「預」，「事」，「之」の 3 字種について判別可能な特徴がみられる．このうち，「預」，「事」は，標題文字列の各々先頭，末端文字である．



原画像　　線素方向　　標本点抽出

図 4.7: 古文書文字の特徴抽出

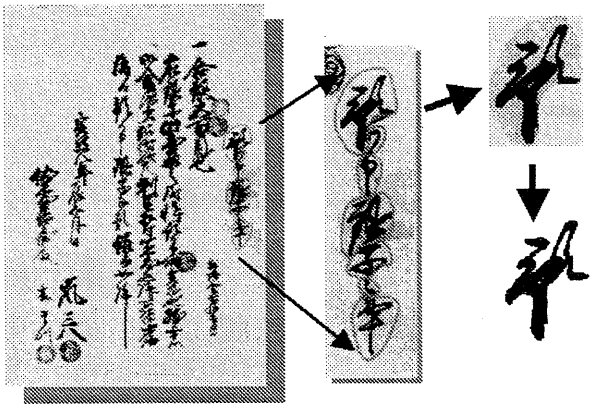


図 4.8: 文字切り出しとパターン辞書作成

4.3.2 正規化による類似性

前項で扱った文字パターンを用いて，前述した文字認識プロセスの正規化過程での文字変形操作により，まったく異なる字種との類似性について調べる．

正規化の操作は，図 4.6 に基づくサイズ変換操作をおこない，図 4.15 では，「事」，「申」，「ヶ」が各々矩形で囲まれた文字に変換される．実験は，正規化後に重ね合わせ法を用いて類似度が高い文字パターンを抽出している．「事」は，原字形の特徴である縦長の特徴を失っている．図 4.16 では，「上」，「事」の類似性がたかく，推定したとおりまったく異なる字形との類似性がたかくなっていることが判る．

W/H 比 r $<$ 1.0 の場合の「払」，「事」，及び「覚」，「定」の類似度がたかい（図 4.17）．W/H 比が r $\fallingdotseq$ 1.0 の場合，「養」，「券」，「合」の 3 字種の類似度がたい．こうした正規化による字形変化は，文字認識過程であきらかに誤認識となる結果を生み出す．古文書文字認識プロセスを検討するうえで，重要な問題のひとつとして検討する必要がある．

4.4 古文書文字認識（HCR）プロセスの検討

これまでの HCR における検討では，文字の共起関係や隣接条件をまったく考慮せずに進めてきた．人間のおこなう文字認識では，前述したとおり文字や語の前後関係，共起関係，及びそれらの背景などの知識を持って解読される．したがって，古文書文字認識においても当然，これらの仕組みを反映させねばならない．n-gram による文

図 4.9: 文字パターンと翻刻文字

字種	ファイル名	W	H	位置	前後文字列
預,	f0001#01.pbm,	95,	135,	f0001, 01,	預り申銀子之事
り,	f0001#02.pbm,	46,	41,	f0001, 02,	預り申銀子之事
申,	f0001#03.pbm,	29,	63,	f0001, 03,	預り申銀子之事
銀,	f0001#04.pbm,	77,	102,	f0001, 04,	預り申銀子之事
子,	f0001#05.pbm,	64,	57,	f0001, 05,	預り申銀子之事
之,	f0001#06.pbm,	31,	26,	f0001, 06,	預り申銀子之事
事,	f0001#07.pbm,	66,	105,	f0001, 07,	預り申銀子之事

図 4.10: 文字パターン属性情報

字の共起関係の検討を本標題文字の認識にも導入し，実験している．

また，前項で述べた文字認識過程で行われる正規化によって本来の文字パターンがもつ属性が失われる．これを改善した文字認識プロセスを提起するとともに n-gram をも併用した認識実験について示す．

4.4.1 n-gram による隣接文字の推定

n-gram とは，文字列の n 文字が隣接して生じる共起関係である．標題文字の認識では，2-gram（2 文字）により，ある文字の後に出現する文字を推定して，これを文字認識プロセスに組み込むことにする．

表 4.18 は，横（行）方向の文字が第 1 字目で，縦（桁）方向の文字が第 2 字目である．例えば，左上端から「＜空白＞」（第 1 字目）から「＜空白＞」（第 2 字目）の頻度が 951 である．続いて，「＜空白＞」から「預」の頻度が 258 である．

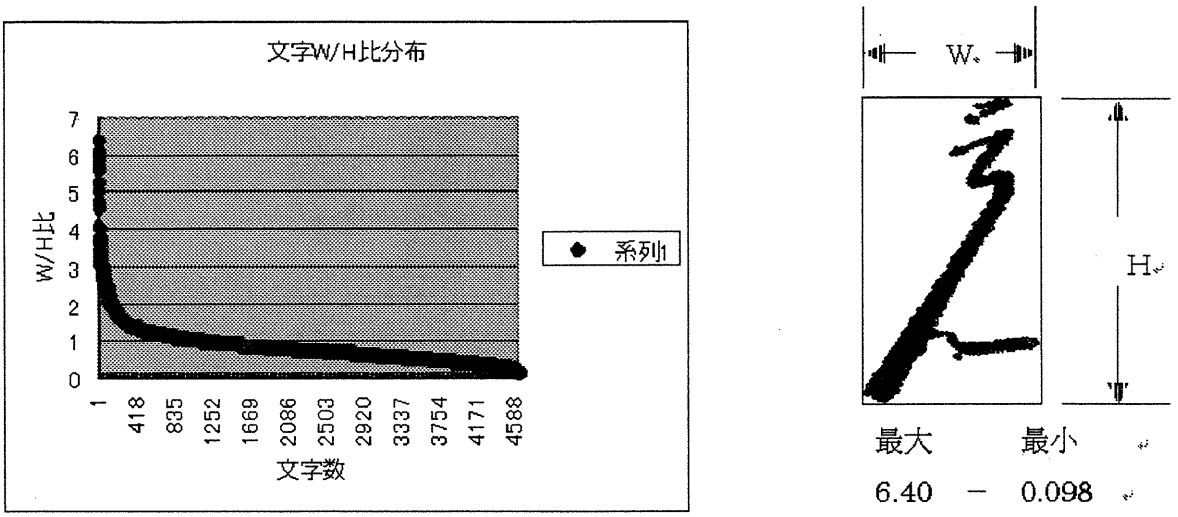


図 4.11: 文字パターンのサイズ特徴

字種	最大値	最小値	平均値	頻度	総字数比
	W/H比			出現文字	
之	5.275862	0.2	0.777284	533	0.115318
事	1.777778	0.098196	0.421203	515	0.111424
申	2	0.194444	0.52975	300	0.064907
預	2.35	0.29703	0.835345	254	0.054955
り	2.09434	0.254902	0.806742	246	0.053224
子	1.743243	0.295566	0.82198	245	0.053007
覚	1.016667	0.316602	0.55079	208	0.045002
金	1.594059	0.23913	0.882304	195	0.04219

■ 総文字数 4,622　　文字種 193

図 4.12: 文字パターンのサイズ特徴

4.4.2 あらたな古文書文字認識プロセスの検討

従来の文字認識過程では，前述したように (1) 切り出しから認識までが順次処理される，(2) 辞書への正規化では失われる情報がある，(3) 文字サイズ，意味カテゴリーなどをパラメタにした辞書検索をおこなっていない，(4) 通常は，認識過程の終了後の後処理で整合性がチェックされる．

こうした，従来型の認識プロセスにおいて，人間の文字認識プロセスに近いモデル化が可能かどうかを検討する．具体的には，

1. 各文字パターンのサイズなどの特徴が失われない方法
2. 辞書検索時にサイズ等のパラメタが指定できる
3. 後処理から認識へバックトラックする機能
4. 文字切り出しと認識の同時処理が行われる方法

などを検討する必要がある．

以下に示す文字認識の実験では，上記の 1，4 について実現する．正規化は，認識しようとする対象画像に対して，文字パターン辞書から取り出されたパターンを対象画像のサイズに一致するように変換することである．従

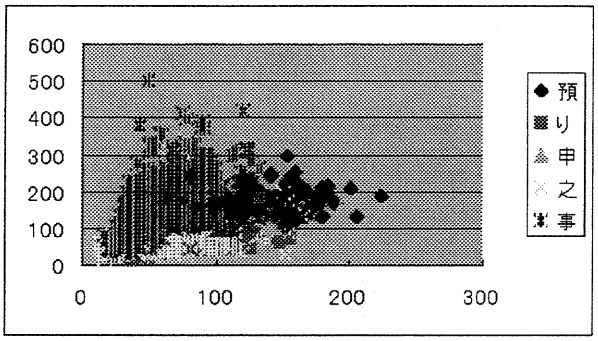


図 4.13: W/H による散布図

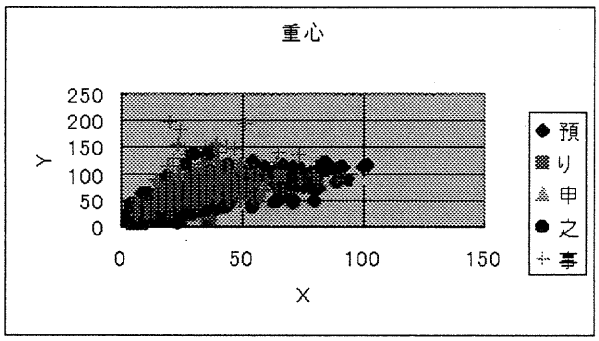


図 4.14: 重心による散布図

来の認識プロセスとはまったく逆の発想で検討した.

まず, 2-gram を用いた切り出し, 及び認識プロセスについて検討する.

1. 標題の先頭文字に出現する文字カテゴリーに含まれる 1 文字パターン (図 4.20 : 右側文字パターン) を辞書から取り出す. その際には, サイズ等の情報が有用であるが, 本実験では使っていない. 文字幅 c は, 文字パターンから得られる範囲 a 内の最大字幅とする.
2. つぎに対象画像の文字幅 c に, 辞書から取り出した文字パターン (図 4.21:字幅 : d) を幅 c に変換する. すなわち, 正規化する.
3. つぎにマッチングに移行する. マッチングは重ね合わせ法によるが, 隣接文字の「侵入」や「連結」を切出すためにマッチングをおこなう範囲を限定しなければならない. このために, マスク処理 (図 4.22) をおこなう.
4. 対象画像上での探索範囲は, 概ね経験則から文字パターンのたかさの 2 倍としている (図 4.23).
5. マッチングにより, 両パターンの距離が一定のしきい値以下になったとき, 一致したと見なす.
6. 一致したパターンで対象画像のパターンを消去し, これがつぎの対象画像となる (図 4.24).

以上があらたな試みの認識プロセスの概要である. この実験結果から, 2-gram を用いて切り出し・認識を行った場合, 約 90 %の認識率を得た. この方式は, 従来の人間の動作に比較してより近いのではないかと考えている.

4.5 おわりに

従来型の文字認識モデルに従う古文書文字認識をおこなう場合に, 文字パターン辞書の特にサイズに一致させる操作, すなわち正規化において, すくなくともまったく異なる字形との類似度がたかくなる場合がある. これ

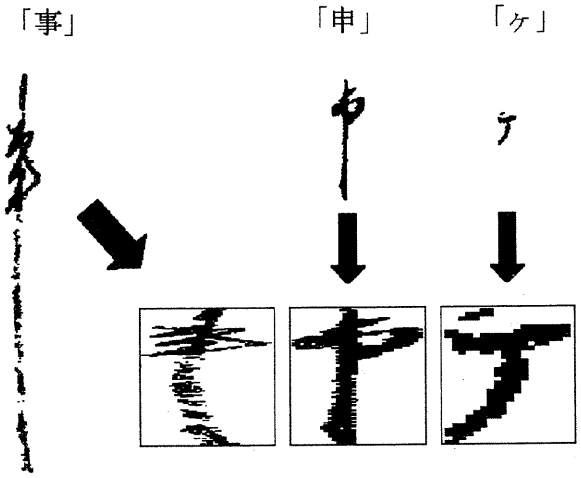


図 4.15: 正規化による字形変化　その 1

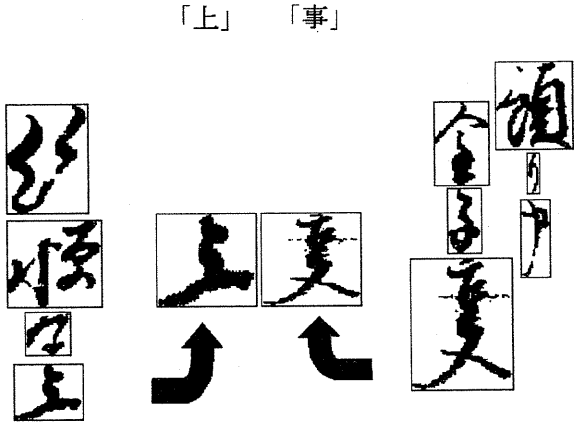


図 4.16: 正規化による字形変化　その 2

は，認識過程での認識率に影響すると推定され，今後も引き続いての検討が必要である．また，従来型の文字認識プロセスとは異なる古文書文字認識プロセスについて検討し，その実験結果について示した．本実験では極めて限られた標題，字数の範囲での実験であり，多くの問題点を含む．たとえば，文字パターン辞書から選ばれたパターンの W/H 比が $r > 1.0$ で，極端に r 値が大きい場合の切り出し手法や，また隣接文字の先に出現した文字で認識に失敗した場合，引き続く文字の認識にも影響する．さらに，対象画像の文字サイズがちいさい場合に類似性がたかくなることなどが問題である．

また，従来型の文字認識プロセスにおいても認識過程で，文字サイズ等の属性を上手に活用する手法を工夫しなければならない．これは文字パターン辞書と検索・参照，及び辞書構築の研究でもある．

今後，さらに文字パターンの特徴を調査するとともに，前項でおこなった実験でパターン数を増やした実験を計画している．

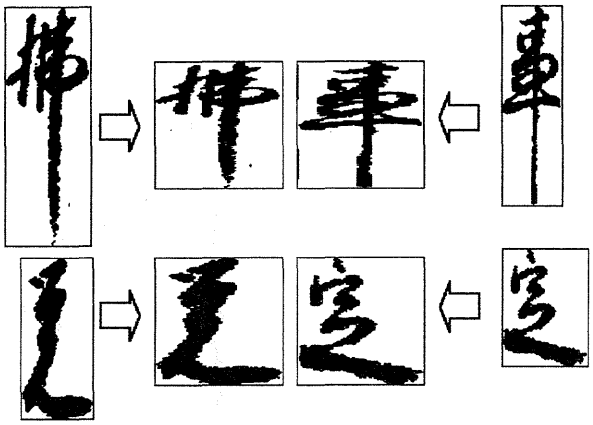


図 4.17: W/H 比 r < 1.0 の字形変化

		預	り	申	銀	子	之	事	役	者	手	附	證	文	住	一	札	相	借	用	金	覚	歌
	951	258	0	0	0	0	0	0	3	0	12	2	5	1	0	33	0	1	71	0	3	204	19
預	0	0	276	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
り	0	1	0	251	0	0	2	0	0	0	0	0	4	0	0	5	0	0	0	0	1	0	0
申	1	0	0	0	70	1	0	1	0	0	18	0	2	0	0	21	0	0	0	0	184	0	0
銀	1	0	0	0	0	70	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
子	7	0	0	0	0	0	242	1	0	0	0	0	2	0	0	0	0	0	1	0	0	0	0
之	22	0	1	0	0	0	0	521	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
事	525	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
役	0	0	0	0	0	0	0	0	0	22	0	0	0	0	0	0	0	0	0	0	0	0	0
者	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19	0	0	1	0	0	0	0	0
手	0	0	0	0	0	0	0	0	0	0	0	37	0	0	0	0	0	0	0	0	0	0	0
附	2	0	0	0	0	0	0	0	0	0	0	0	33	0	0	0	0	0	0	0	3	0	0
證	2	0	0	0	0	0	0	0	0	0	0	0	1	89	0	1	0	0	0	0	0	0	0
文	12	0	0	0	0	0	73	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0
住	0	0	0	3	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
一	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	142	0	0	0	0	0	0
札	45	0	0	0	0	0	99	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
相	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0
借	0	0	0	1	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	17	0	0	0

標題総数: 908　総文字種(文字パターン数): 196　総文字数: 5,628 字(空白 878 字を含む)

図 4.18: 2-gram による標題文字の出現頻度

預	り	申	ケ																				
285	276	8	1																				
り	申	渡	一	證	之	受	預	金	茶														
274	251	7	5	4	2	2	1	1	1														
申	金	銀	一	手	家	證	頼		子	事	年	約	畑										
308	184	70	21	18	5	2	2	1	1	1	1	1	1										
銀	子	請	之		諸																		
77	70	3	2	1	1																		
子	之		證	事	借	年																	
254	242	7	2	1	1	1																	
之	事		内	り	通																		
548	521	22	3	1	1																		
事		請																					
527	525	2																					

図 4.19: 「預り申銀子之事」の 2-gram 表

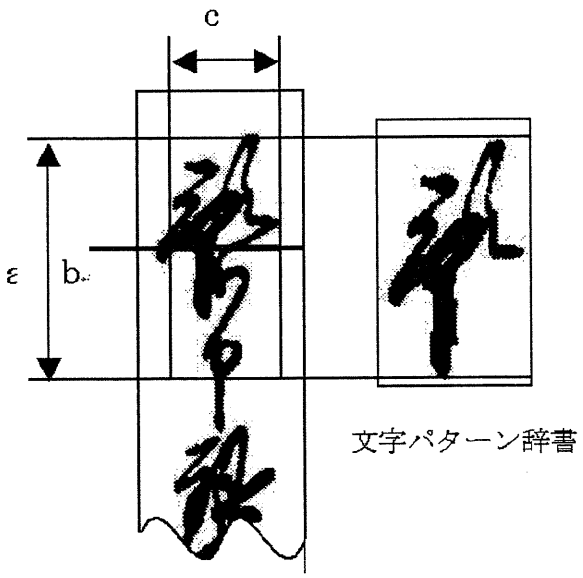


図 4.20: 対象画像の字幅検出

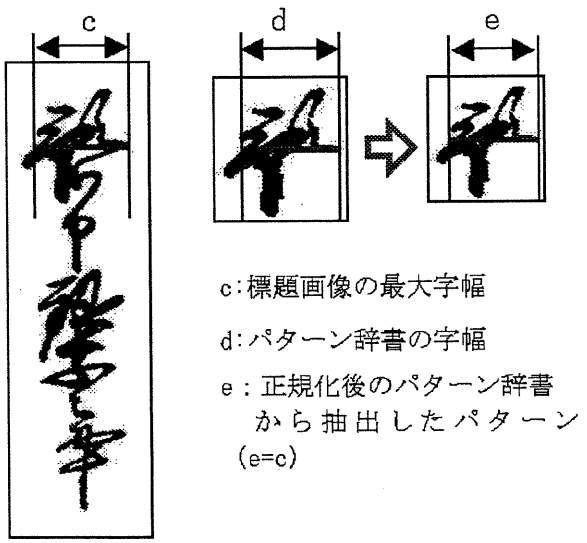


図 4.21: 対象画像（文字）への正規化

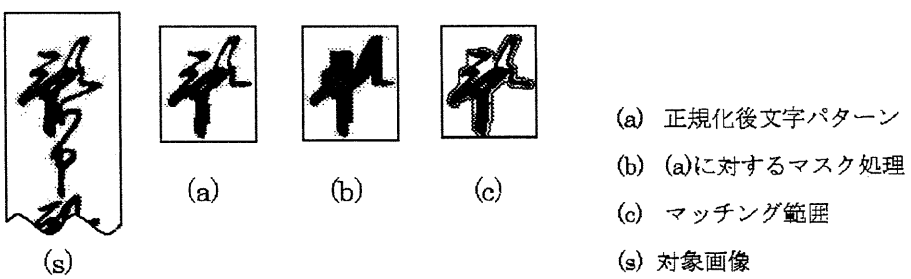


図 4.22: 認識領域のマスク

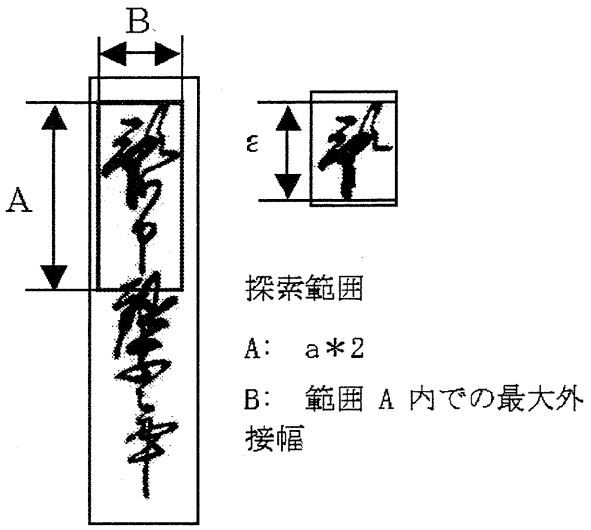


図 4.23: 探索範囲の決定

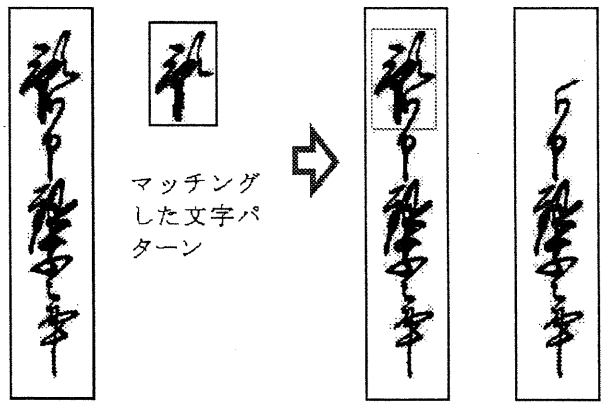


図 4.24: マッチングと文字消去

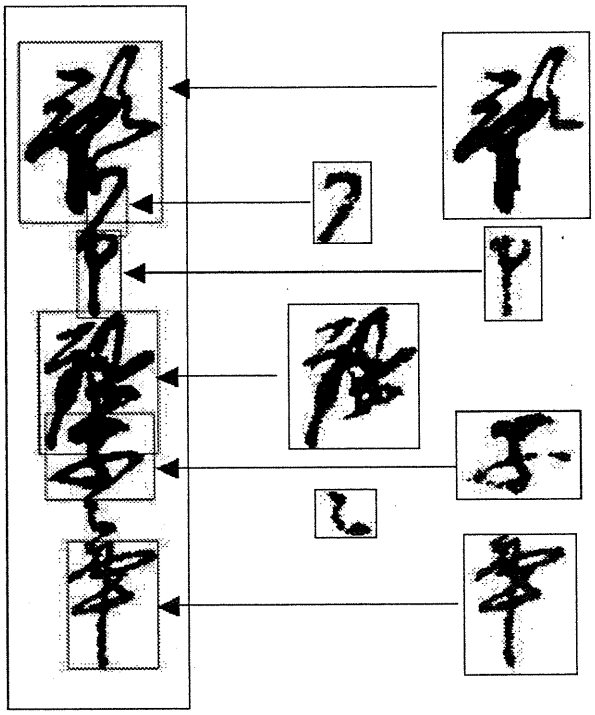
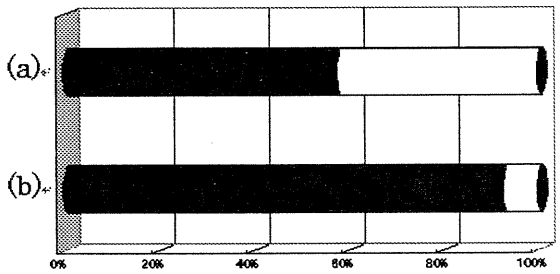


図 4.25: 切り出し・認識結果例



■ 正しい認識

□ 誤認識

(a)2-gram 未使用　認識率 57.7%

(b)2-gram 使用　　認識率 90.7%

総文字パターン数：97

図 4.26: 切り出し・認識実験結果

第5章

古文書文字認識の実験

5.1 まえがき

近年，パターンの統計的性質を用いた文字認識技術の研究が盛んに行われている．その中で，ベイズ識別やマハラノビス距離を用いた認識システムの有効性が確認されている．一般にそれらの手法は，パターンの分布が正規分布をしていると仮定し，学習サンプルからその分布の推定を行っている．パターンの分布の推定を行う際，十分な学習サンプル数を確保できる場合は，非常に高い認識精度を得ることができる．しかし，学習サンプル数が少ない場合や字種間で異なる場合は，共分散行列などの認識に必要なパラメータを高い精度で推定することができず，認識精度の低下が生じてしまう問題点がある．図5.1に示されるような古文書文字の認識問題の場合，同一字種であってもくずしや書風により文字の形状が異なり，それらの分布が複数のクラスによって構成されていると考えれる．そのため，上記した統計的手法のように学習対象の分布形状を仮定しなければならない場合，その仮定が認識対象に対し妥当なものでなければ，高い認識精度は期待できず，より柔軟な認識手法の確立が必要であると言える．

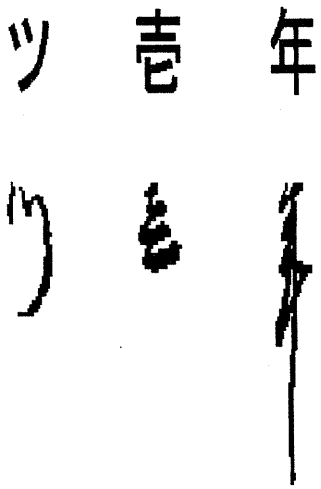


図5.1: 古文書文字の例

そのような認識問題に有効であると考えられる手法に，ニューラルネットワークを用いた文字認識手法がある．ニューラルネットワークは，その柔軟な情報処理と高い汎化能力により，高い認識精度が期待できる．しかし，そ

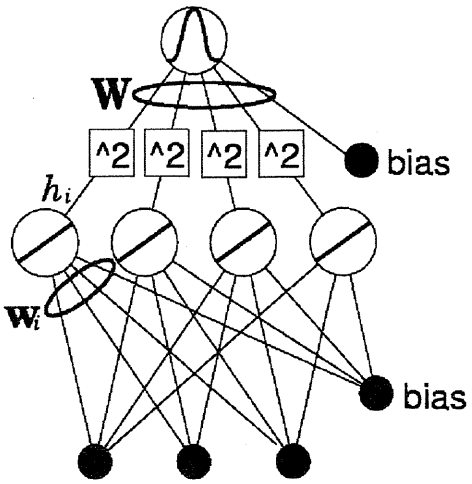


図 5.2: ネットワークモデル

の学習は，学習サンプルを繰り返し投入しながら学習する必要があり，統計的な手法と比較して学習の計算量が膨大になる欠点があり，十分な学習が行うことが不可能であった．しかし，最近の計算機の性能向上に伴い，従来，学習が困難であった認識問題に対しても十分な学習を行うことが可能になり，統計的手法に迫る認識精度を確保できるようになって来ている．

主に，文字認識に用いられるニューラルネットワークモデルとして，MLP(Multi Layerd Perceptron)[15], LVQ(Learning Vector Quantization)[16], RBF(Radial Basis Function)[17] などが挙げられる．一般に，最も高い認識精度が得られているモデルは MLP である．MLP は教師あり学習である Back Propagation によって学習を行う．教師あり学習を行うことにより，一つのネットワークが複数の字種を学習の対象とすることが可能となる．このことは，各字種の分布を推定するだけでなく，識別に必要な字種間の差異を学習することが可能であると考えられ，特に，形状の似た類似字種に対し高い認識精度が期待できる．そこで本稿では，ニューラルネットワークを古文書個別文字認識に適用し，その認識性能を統計的手法と比較して報告する．

5.2 ニューラルネットワークのモデルと動作

5.2.1 ネットワークモデル

認識実験に用いるネットワークモデルを図 5.2 に示す．ネットワークは入力層を含めて 3 層構造で，通常の MLP とは異なり，隠れ層と出力層の間に伝達する信号を自乗する機能を持つ自乗結合を導入している．隠れ層ニューロンは線形の活性化関数を用い，出力層ニューロンはガウス型の活性化関数を用いる．

5.2.2 ネットワークの出力と学習

順伝搬

入力ベクトルを $\mathbf{x}$，隠れ層ニューロン i の重みベクトルを $\mathbf{w}_i$，出力層ニューロンの重みベクトルを $\mathbf{W}$ とすると，ネットワークの出力 O は以下の式で定義される．

$$h_i = \mathbf{x} \cdot \mathbf{w}_i + \theta_i \tag{5.1}$$
$$H_i = h_i^2 \tag{5.2}$$
$$O = exp(-\mathbf{W} \cdot \mathbf{H} + \theta) \tag{5.3}$$

ここで，h_i は隠れ層ニューロン i の出力，$\mathbf{H}$ は h_i を自乗した H_i を成分に持つベクトル，θ_i，θ は bias ニューロンとそれに対する重みにより決定される値である．

逆伝搬

ネットワークの学習は，学習ベクトルとそれに対する教師信号の対を $(\mathbf{x}^m, T^m), m = 1, \cdots, M(M$ は定数$)$ とした場合，式 (5.4) で定義される誤差に対し Back Propagation を適用して行う．

$$E = \frac{1}{2}\sum_{m=1}^{M}(T^m - O^m)^2 \tag{5.4}$$

自乗結合を導入したことにより学習則は次式のようになる．

$$\begin{aligned}\frac{dW_i}{dt} &= -\alpha\frac{\partial E}{\partial W_i} \\ &= -\alpha\sum_{m=1}^{M}(T^m - O^m)O^m h_i^{m2}\end{aligned} \tag{5.5}$$

$$\begin{aligned}\frac{dw_{ij}}{dt} &= -\alpha\frac{\partial E}{\partial w_{ij}} \\ &= -2\alpha\sum_{m=1}^{M}(T^m - O^m)O^m W_i h_i^m x_i^m\end{aligned} \tag{5.6}$$

ここで，W_i と w_{ij} は，それぞれ，$\mathbf{W}$ と $\mathbf{w}_i$ の各成分である．また，式 (5.1)，(5.3) の θ_i，θ は重み w_{ij}，W_i の一成分として表現した．

5.2.3 Weight Decay

ニューラルネットワークの汎化能力を向上させる手法として，式 (5.7) に示される Weight Decay がある．Weight Decay は，各ニューロンの持つ重み $\mathbf{w}$ に式 (5.7) を適用するアルゴリズムである．これは，重みベクトルが長くなると，ニューロンの活性化関数の傾きが急峻になり，未知入力の変動に過敏反応してしまうことを防ぐ効果がある．

$$\mathbf{w}_{t+1} = (1 - \beta)\mathbf{w}_t \tag{5.7}$$

β は崩壊のパラメータで $\beta << 1$ である．

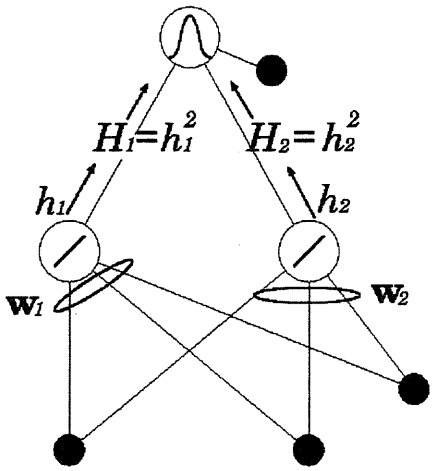


図 5.3: 2 次元の楕円の学習:ネットワーク構成

2 次元の楕円の学習

このモデルがどのような分布を表現し得るかを，2 次元空間での楕円の学習を行うことで確認する．学習に用いるネットワーク構成は図 5.3 に示されるように 2 入力 1 出力で隠れ層ニューロン数は 2 とした．学習データは一様な分布の中からランダムに選出し，

$$(x-3)^2+(x-3)y+y^2<0.2 \tag{5.8}$$

を満たす場合は教師信号に 1.0 を，それ以外は 0.0 を与え学習に用いた．

図 5.4 に学習データとネットワークの出力の等高線を示す．図から分かるように，学習対象としている 2 次元空間を構成する軸に対し，角度を持った楕円の長軸と短軸を学習出来ていることが分かる．

5.3 認識システムの概要

実験に用いる認識システムの概要を図 5.5 に示す．

5.3.1 前処理

特徴抽出の前処理は，ノイズ除去としての孤立点除去と，文字の大きさの正規化を行う．特徴抽出で用いるイメージサイズが 64 ドット× 64 ドットであるため，入力イメージの幅と高さの大きい方をが 64 ドットになるような倍率で，入力イメージの縦と横の比率を保つように正規化を行う．

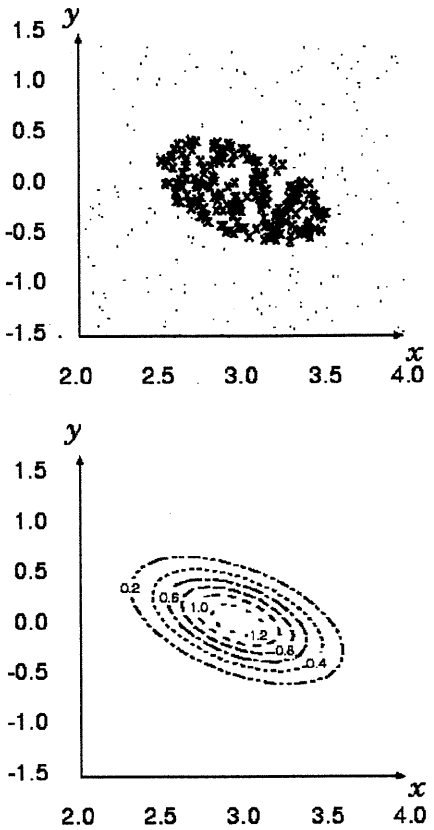


図 5.4: 2 次元の楕円の学習:学習データとネットワークの出力の等高線

5.3.2 特徴抽出

特徴量として改良型方向線素特徴量 [18] を用いる．改良型方向線素特徴量は 196 次元で構成される．抽出アルゴリズムは，前処理を施されたイメージに対し，輪郭線抽出・線素処理化を行う．線素としては縦（｜），横（—）右上斜め（／），左上斜め（＼）の 4 種類を割り当てる．次に，これを 8 ドット ×8 ドットの正方領域に分割し，その隣り合う 4 個ずつを一つの小領域とする．全部で 49 個の小領域となる．各小領域毎に線素の数を重み付きで数えることで特徴量とする．小領域 49 個 ×4 種類の線素のため，196 次元のベクトルが得られる (図 5.6)．

5.3.3 大分類部

大分類部として，各字種の平均ベクトルを用いたパターンマッチングを用いる．距離尺度としては，ユークリッド距離を用いる．大分類部により，ある程度細分類部に入力する字種を削減することにより，誤認識が生じる確率を抑えることができる．

5.3.4 細分類部

認識システムの細分類部として **5.2.** に示したニューラルネットワークを用いる．細分類部の構成は，一つのネットワークが特定の一字種のみに発火するように割り当てられたモジュラー型のネットワーク構成とする．個々のモジュールは，割り当てられた字種に対しては発火し，それ以外の字種に対しては発火を抑制するように学習を

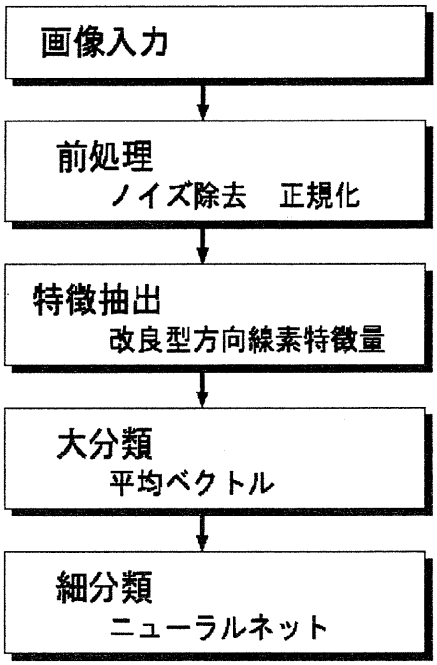


図 5.5: 認識システムの概要

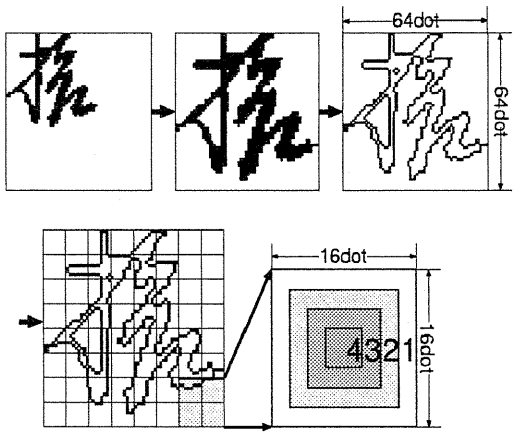


図 5.6: 方向線素特徴量

行う.

5.4　古文書文字認識

古文書文字の認識実験を行う．**5.3.** で述べた認識システムの細分類部に，**5.2.** のニューラルネットワークを用いた場合と，統計的手法の一つである改良型マハラノビス距離を用いた場合とを比較する．

5.4.1　使用データ

実験に用いるデータは，「宗門改帳」古文書画像データベースに登録されている古文書画像から，川口ら [19] によって収集された 16 字種 (ツ，一，二，三，四，五，六，七，八，九，十，壱，弐，年，拾，廿) とする．各字種のサンプル数は，「廿」が 66 個で，その他の字種は 200 個である．学習には，「廿」以外は 80 個，「廿」は 33 個の

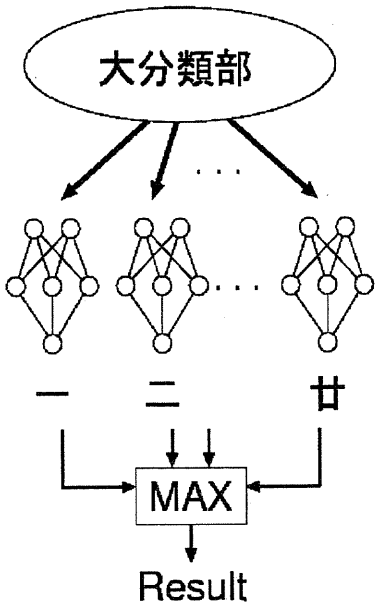


図 5.7: ニューラルネットワークの構成

サンプルを用いる.

5.4.2 ニューラルネットワークの構成

入力層，隠れ層，出力層の 3 層構成で，ニューロン数は，それぞれ，196，30，1 とした．図 5.7 に示されるように，各字種に一つの MLP モジュールが割り当て，最大の出力を得たモジュールに割り当てられた字種を認識結果とする．各モジュールは，割り当てられた字種に対し教師信号を 1.0，それ以外の字種に対し教師信号を 0.0 として学習を行う．学習率が 0.00001，教師信号との誤差絶対値の平均が 0.02 以下，または学習回数が 100 回を越えるまで学習を行う．Weight Decay の崩壊パラメータ β は，0.0(Weight Decay 無し) と 0.000001 とした．

5.4.3 改良型マハラノビス距離

改良型マハラノビス距離 [20] は式 (5.9) で定義される．共分散行列の固有値にバイアスを加えることによって，小さい固有値の方距離に大きな影響を与えてしまうことを防いでいる．

$$D_m(\mathbf{x}, \mathbf{u}^i) = \sum_{j=1}^{k} \left(\frac{1}{\lambda_j + b}\right) \left((\mathbf{x} - \mathbf{u}^i)^t \mathbf{e}_j\right)^2 \tag{5.9}$$

ここで，b，$\mathbf{x}$，$\mathbf{u}$ は，それぞれバイアスと入力ベクトル，標準パターンベクトルを表し，$\mathbf{e}_j$ は，固有値 λ_j に対する固有ベクトルで，$\lambda_j \leq \lambda_{j+1}$ である．

共分散行列から算出できる固有ベクトル数は，学習サンプル数によって決定され，サンプル数が 80 個の場合 79 個，33 個の場合 32 個となる．認識時には，「廿」以外の字種の標準パターンからの距離を求める時は $k = 79$，「廿」の標準パターンからの距離を求める時は $k = 32$ とした．計算に用いる次元数が小さい方が距離が小さくなるため，実験では，D_m/k のように正規化された距離を用いて認識を行う．

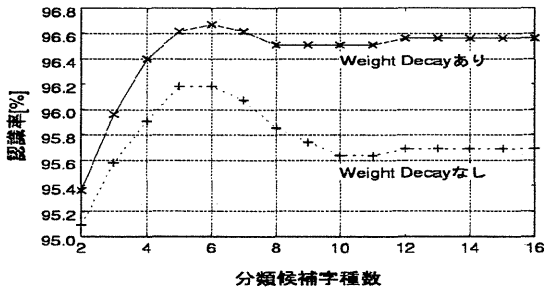


図 5.8: ニューラルネットワークを用いた認識率

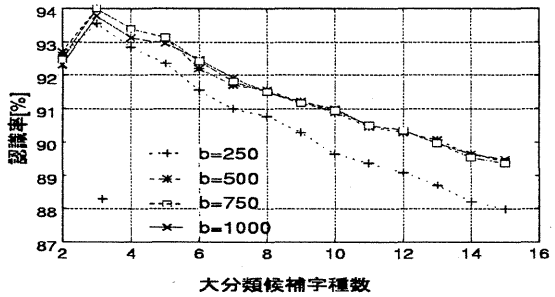


図 5.9: 改良型マハラノビス距離を用いた認識率

5.4.4 認識結果

図 5.8, 5.9 に，ニューラルネットワークを用いた認識率と改良型マハラノビス距離を用いた認識率をそれぞれ示す．図から分かるように，ニューラルネットワークの Weight Decay を適用したものが最も高い認識率の 96.67%が得られている．参考までに，「廿」を除いたニューラルネットワークと改良型マハラノビス距離の認識率はそれぞれ，97.05%と 94.22%である．

表 5.1, 5.2 に，それぞれ，Weight Decay を適用したニューラルネットワークと改良型マハラノビス距離の誤認識を起こした字種の内訳を示す．これらは，それぞれの手法で最も高い認識率を得られた大分類候補字種数を用いた場合である．ニューラルネットワークを用いた場合では，サンプル数の少い「廿」を除いて全て 90%以上の認識率を得られている．しかし，「廿」が他の比べ 20%程低い認識率となってしまった．「廿」の学習サンプル数は，他の字種の半分以下であったため，そのことが認識精度低下の原因になったと考えられる．ニューラルネットワークの認識精度とカテゴリ間の学習サンプル数の違いがどのとような関係にあるか，調査が必要である．

表 5.1: 各字種の認識率 (ニューラルネットワークを用いた場合)

字種	正読数	誤読数	認識率 [%]
ツ	117	3	97.50
一	120	0	100.00
二	116	4	96.67
三	120	0	100.00
四	117	3	97.50
五	114	6	95.00
六	111	9	92.50
七	115	5	95.83
八	116	4	96.67
九	117	3	97.50
十	117	3	97.50
壱	118	2	98.33
弐	114	6	95.00
年	116	4	96.67
拾	119	1	99.17
廿	25	8	75.76
合計	1772	61	96.67

表 5.2: 各字種の認識率 (改良型マハラノビス距離を用いた場合)

字種	正読数	誤読数	認識率 [%]
ツ	117	3	97.50
一	116	4	96.67
二	110	10	91.67
三	116	4	96.67
四	118	2	98.33
五	111	9	92.50
六	109	11	90.83
七	106	14	83.33
八	106	14	83.33
九	113	7	94.17
十	117	3	97.50
壱	115	5	95.83
弐	112	8	93.33
年	113	7	95.17
拾	117	3	97.50
廿	27	6	81.82
合計	1723	110	94.00

5.5 まとめ

本稿では，古文書文字のようなくずしなどの変形の多い認識問題に対し，柔軟な情報処理が可能なニューラルネットワークが有効であると考え，統計的手法と比較してその認識精度を実験的に求めた．統計的手法の一つである改良型マハラノビス距離と比較して2.6%程高い認識率が得られたが，字種間の学習サンプル数に差があり，学習サンプル数が少い字種の認識精度が極端に低くなってしまうことが明らかになった．古文書文字の場合，認識対象とする字種の十分な数のサンプルを収集することは困難であると考えられ，少い学習サンプル数や字種間に偏りがある場合に対しても，高い認識精度を実現し得るネットワークアーキテクチャや少い学習サンプルから認識精度の向上を可能にする学習サンプルの生成手法の検討が今後の課題であると言える．

第 6 章

文字切り出しを前提としない古文書標題認識

6.1 はじめに

古文書翻刻支援システムの開発では，古文書がくずし字やつづけ字で書かれることから，従来の文字認識技術を用いることは難しい．これは認識を行うために，あらかじめ文字列からの文字切り出しを前提としているためである．そこで本研究では，従来の文字認識過程とは異なり，文字認識の対象となる標題画像の射影ヒストグラムから推定した探索範囲に対して，文字パターン辞書から取り出した文字パターンを探索範囲の文字幅で正規化しテンプレートとしてマッチングを行う，切り出しを前提としない認識手法について提案し，その有効性について検討する．

6.2 文字切り出しを前提としない文字認識手法

6.2.1 従来の文字認識過程

従来用いられてきた一般的な認識過程を図 6.1 に示す．まず認識対象となる文字列に対して，ノイズ除去，スムージングなどの前処理を行う．次に文字列から各文字や語単位で文字を切出す．そして切り出した文字を辞書の文字パターンに合せるように正規化し，認識を行う．

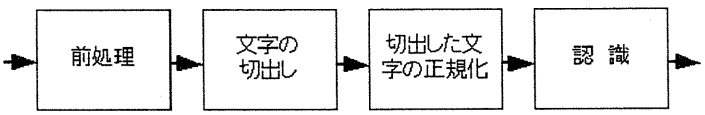


図 6.1: 従来の文字認識過程

従来の認識過程を古文書に適用させた場合，各文字や語が適切に切出せるかが問題になる．なぜなら図 6.2 のようなくずし字やつづけ字が多い文字列から切り出しでは，良い結果が得られていない．

そのため切り出した文字には，上下文字や他行からの接触や侵入などの影響によるノイズが含まれていることが多い．これらのノイズを除去できなければ，認識精度の低下につながることが予想される．

6.2.2 本手法の文字認識過程

本手法では，まず認識対象となる文字列に対して，用意した文字パターン群（以下文字パターン辞書という）とのマッチングを行う範囲（以下探索範囲という）を設定する．次に探索範囲内の文字とマッチングを行うために，文字パターン辞書から取り出した文字パターンを，探索範囲内の文字の大きさに合わせるように正規化する．そ

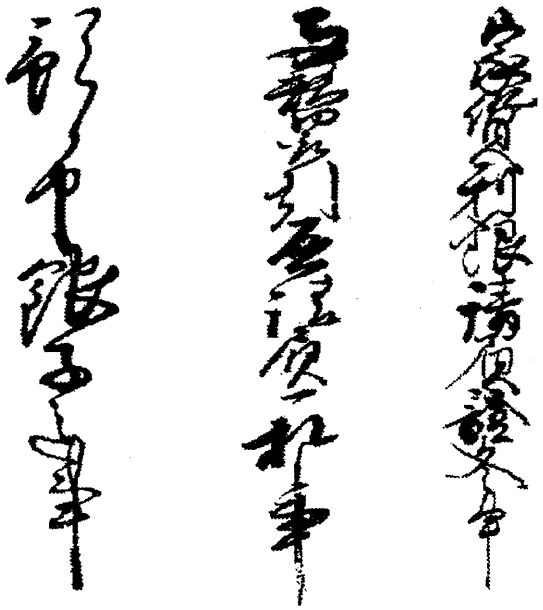


図 6.2: 古文書文字列

して探索範囲内で，文字パターンを左上から右下へと走査させながらマッチングを行う．本手法では認識部の前に文字の切り出し過程を必要としない，つまり前提としていないのが特徴である．また辞書から取り出した文字パターンを，探索範囲内を走査させながらマッチングを行うので，探索範囲に上下文字との接触や，他行からの侵入などのノイズが含まれていてもマッチングの結果に影響を及ぼしにくい．例えば図 6.3 において，探索範囲内には「預」と「り」の 2 文字が存在するが，文字パターン「預」を走査させた場合，探索範囲内の「預」の場所で最大のマッチング結果が得られる．つまりノイズの影響を受けにくいのが分かる．

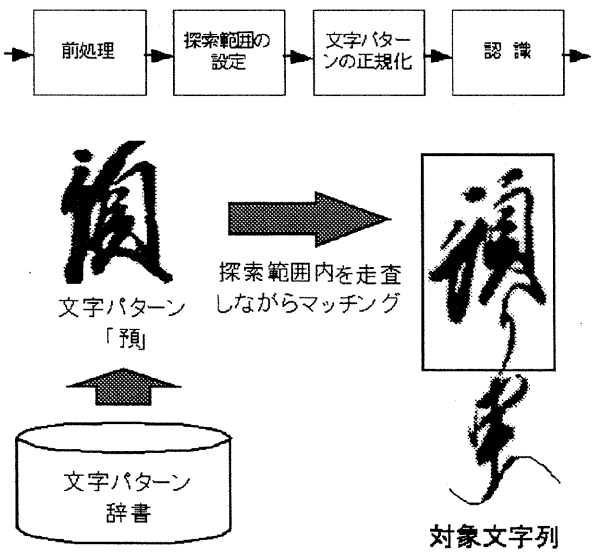


図 6.3: 本手法の文字認識過程

6.3 探索範囲と文字パターン辞書の正規化

6.3.1 ヒストグラム

探索範囲の設定にはヒストグラムを用いる．従来から文字列の特徴を把握するのに，水平方向画素値に基づく射影ヒストグラムが用いられる [14]．しかし毛筆のつづけ字の多い古文書では，ヒストグラムの起伏や切れ目が判断しづらく，特徴を把握しにくい．そこで，最左端の画素から最右端の画素までの距離（文字幅）をヒストグラム化することにより特徴が掴みやすい（図 6.4）．

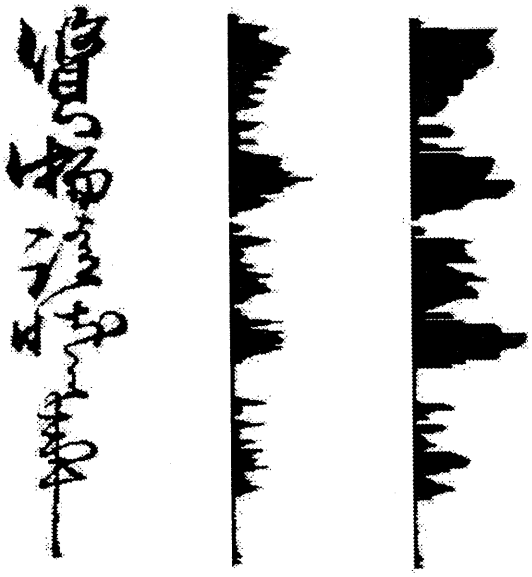


図 6.4: ヒストグラム

6.3.2 探索範囲の設定

まず文字列からストローク幅推定値 [21] を求める．ストローク幅推定値というのは，文字列に含まれる線幅の推定値のことである．次にこの値を閾値とし，ヒストグラムの閾値以下の部分を除去する．これにより，ヒストグラムをいくつかの塊に分割する事が出来る．そして，分割したヒストグラムの上端から下端までの範囲を探索範囲として設定する（図 6.5）．

このとき，上端から下端までの距離が短い場合，つまりあまりに小さく分割されてしまったヒストグラムの塊は，ノイズとして無視する．

図 6.5: 探索範囲の設定

6.3.3 文字パターン辞書の正規化

設定した探索範囲内の文字と，文字パターン辞書の文字の大きさは異なる．そのため，文字パターン辞書の文字を探索範囲の文字の大きさに合うように正規化を行う．

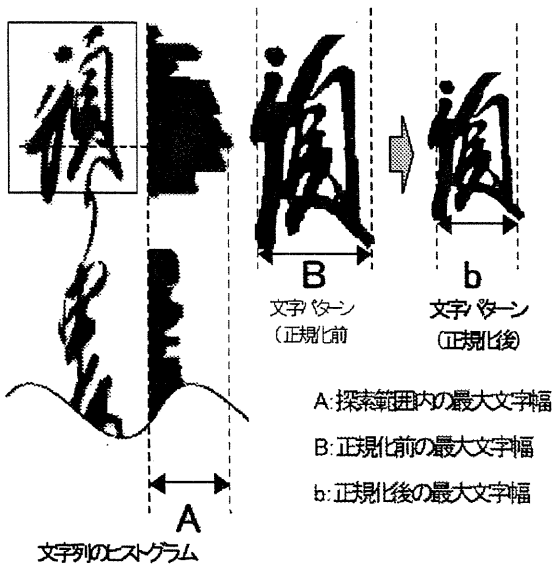


図 6.6: 文字パターン辞書の正規化

まず探索範囲内の文字から最大文字幅を検出する．次に文字パターンに対しても，同様に最大文字幅を求める．そして，探索範囲の最大文字幅と文字パターンの最大文字幅の長さが等しくなるように，文字パターン辞書を拡大，または縮小する（図 6.6）．

6.4 候補文字の抽出実験

6.4.1 実験方法

本手法を用いた候補文字抽出実験行った．実験対象となる文字列は，「伏見屋善衛兵文書」の 200 標題とし，文字パターン辞書として 4420 個の文字パターン（143 文字種）を用意した．ともに「古文書翻刻支援システム開発プロジェクト」のホームページで公開されており，標題画像は「HCD2」，文字パターン辞書は「HCD3」である．

マッチング手法はテンプレートマッチングとし，残差割合の小さい文字から順に，第 10 位まで候補文字として抽出する．そして探索範囲内の文字が，候補文字として抽出できれば正解とした．

6.4.2 実験結果

200 標題に含まれる総文字数 1378 に対して，設定できた探索範囲は 814 である．そしてこの探索範囲を対象とした候補文字の抽出では，59.5 %の累積正解率が得られた（図 6.7）．

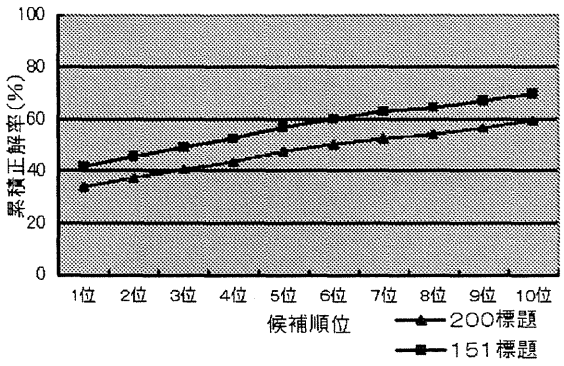


図 6.7: 候補順位別累積正解率

今回用意した標題文字列の中には，文字パターン辞書に存在しない文字や，サンプルの少ない文字が含まれており，その文字が探索範囲に設定される場合があった．そこで「辞書に存在しない文字」または「サンプル数の少ない文字」が，探索範囲として設定された 49 標題を除いた場合の結果についても述べる．これは今回マッチング手法として用いたテンプレートマッチングでは，ある程度のサンプル数が必要なためである．そこで 49 標題を除いた 151 標題を対象とした場合では，候補文字の抽出において 69.7 %という正解率が得られた．

6.4.3 考察

今回の実験では，151 標題を対象とした場合でも 69.7 %という正解率しか得られなかった．これは設定した探索範囲の中に，文字の一部分がはみ出しているものや，ひとつの文字に対して複数の探索範囲を設定してしまったもの，また全く文字を含んでいない探索範囲が存在するために，マッチングの精度が低下してしまったからである．これらの設定に失敗した探索範囲は，図 6.8 のように文字の上側が外れるパターン (a)，文字の下側が外れるパターン (b)，文字の上下両側が外れるパターン (c)，そしてそれ以外のその他のパターン (d) に分類できる．

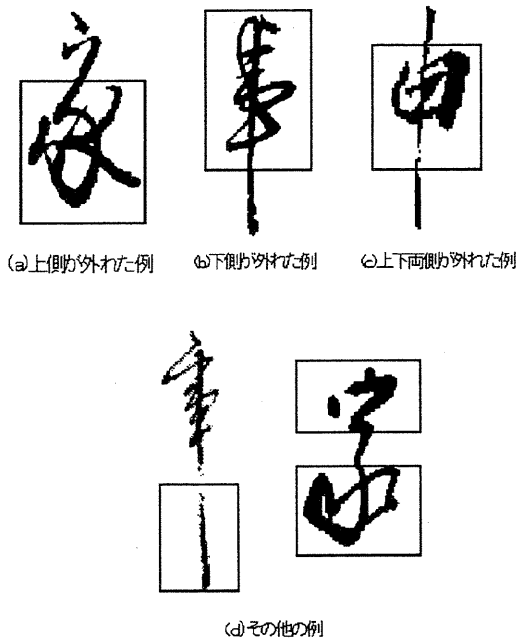
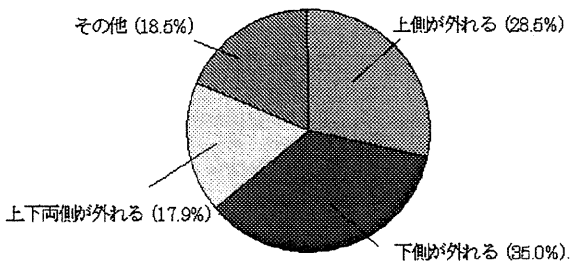
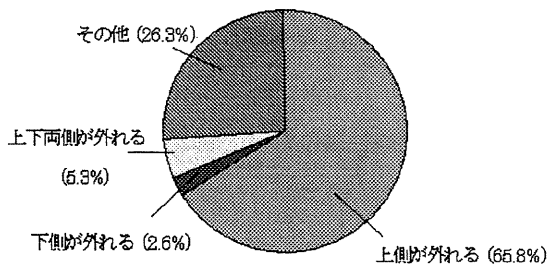


図 6.8: 設定に失敗した探索範囲



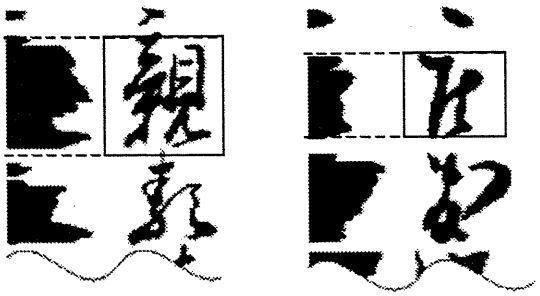
(a) すべての探索範囲



(b) 先頭の探索範囲のみ

図 6.9: 設定に失敗した探索範囲の要因

(a),(b),(c) のパターンは，探索範囲設定において，標題文字列から求めたストローク幅推定値を閾値として用いたため，文字の縦線のみが現れる部分でヒストグラムが分割されてしまうのが原因である．このような例は，「事」や「申」のような文字に起こりやすい．そして (a),(b),(c) のようなパターンは，設定に失敗した探索範囲の 81.4
そこでこの問題を解決するために，あらかじめ辞書内の文字パターンに対して，上下のストローク幅を切除するという前処理を行う．この処理によって，たとえ文字の一部分がはみ出ている探索範囲であっても，候補文字として抽出できるのではないかと考えられる．次に先頭の探索範囲に注目した時，探索範囲設定に失敗した場合は図 6.8(a) のパターンであることが多い（図 6.9(b)）．これは「家」，「永」，「座」，「親」などの書き出しの点が孤立するために，探索範囲の設定に失敗しやすい文字が，先頭文字となる標題がいくつか存在するからである（図 6.10）．そこで一番上の探索範囲に限り，探索範囲を上方に拡張する処理を行う．



「親」　　「座」

図 6.10: 先頭探索範囲設定の失敗例

6.5 探索範囲の拡張と文字パターンに対するストローク切除

6.5.1 先頭探索範囲の拡張

先頭の探索範囲に限り，範囲を上方へ拡張する．探索範囲の上側に，探索範囲の設定時にノイズとみなされたヒストグラムが存在する場合，そのヒストグラムの上端までを，新たな探索範囲として設定する（図 6.11）．

6.5.2 文字パターンに対するストローク切除

辞書内の文字パターンに対して，文字幅のヒストグラムを求める．次にその文字パターンのストローク幅推定値を求め，閾値とする．そしてヒストグラムを上下双方から走査し，はじめて閾値に達する場所までを切除する（図 6.12）．辞書内のすべての文字パターンに対して同様の処理を行う（図 6.13）．

6.5.3 再実験

探索範囲の拡張と，文字パターンに対するストローク切除の前処理を行ったうえで，再度同様の実験を行った．そして前処理を行った場合（処理あり）と，行わなかった場合（処理なし）の実験結果を比較する．

前章の実験では，設定したすべての探索範囲に対して候補文字の抽出を行った．しかし本稿では図 6.7(a),(b),(c) の失敗パターンを対象として，正解率を向上させるために前処理を行った．そこで今回の実験では，(d) のパターンについては候補文字抽出の対象としないこととした．

まず前処理の有効性を確かめるために，図 6.8(a),(b),(c) の失敗パターンのみを対象とした場合の，処理の有無

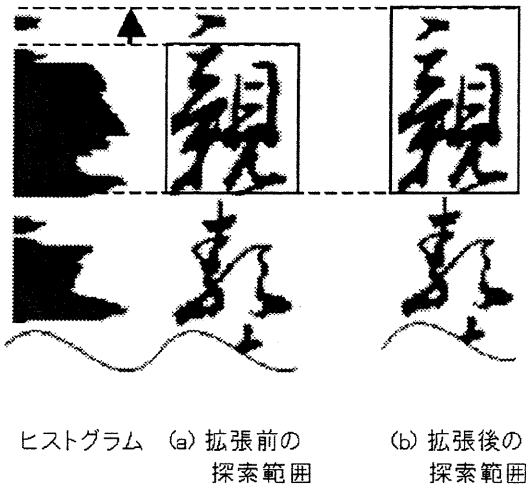


図 6.11: 先頭探索範囲の拡張

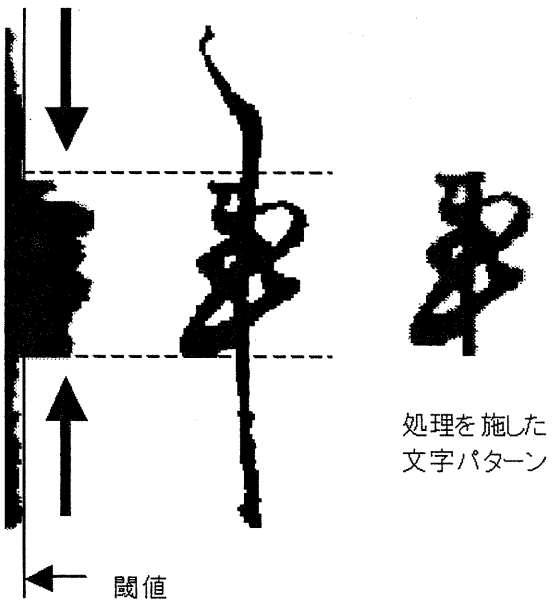


図 6.12: 上下部分のストローク切除

による抽出成功数を表 6.1 に示す.

設定に失敗した探索範囲であっても，200 標題で 44，151 標題で 37 の探索範囲について，新たに正解候補を抽出する事が出来た.

文字パターン辞書に対して前処理を行うことにより，少なからず字形が崩れることになる．これにより，正しく設定された探索範囲の抽出成功数が低下するのではないかと思われる．そこで正しく設定された探索範囲を対象とした場合の処理あり，処理なしの抽出成功数を表 6.2 に示す.

処理の有無でほとんど結果が変わらず，悪影響を与えるどころか，微数ながらも抽出成功数が増加しているのが分かる．これらの結果から，本手法を用いた文字認識において，今回行った前処理が有効である事が分かる.

最後に今回行った実験による，処理の有無による累積抽出成功数を表 6.3 に示す.

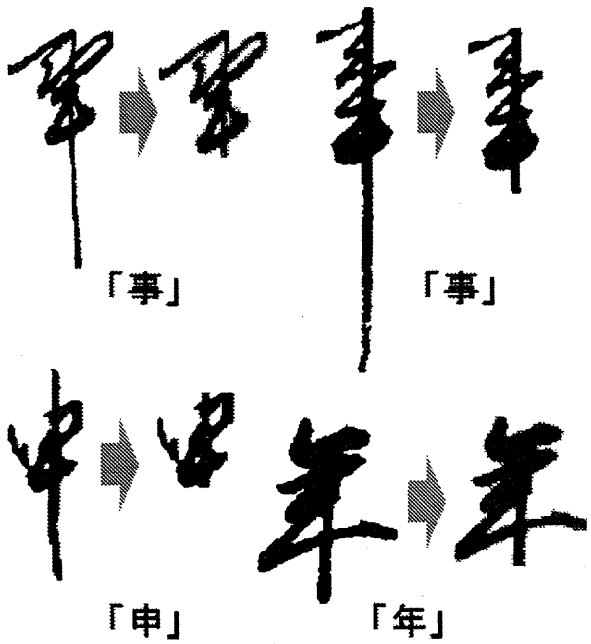


図 6.13: 切除後の文字パターン

表 6.1: 処理の有無による抽出成功数

	対象探索範囲	処理あり	処理なし
200 標題	277	123	167
151 標題	210	106	143

表 6.2: 正しく設定できた探索範囲の抽出成功数

	対象探索範囲	処理あり	処理なし
200 標題	474	361	363
151 標題	340	310	314

表 6.3: 処理の有無による累積抽出成功数

		総文字数	探索範囲数	対象探索範囲数	抽出成功数
200 標題	処理なし	1378	814	751	484
	処理あり				529
151 標題	処理なし	1054	597	550	416
	処理あり				457

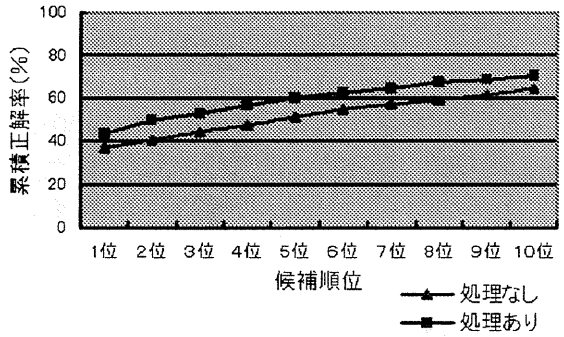


図 6.14: 処理の有無による候補順位別累積正解率（200 標題）

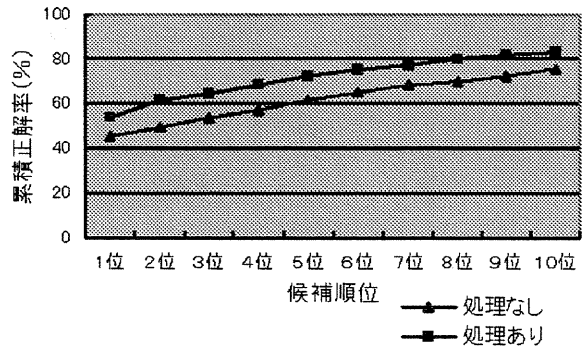


図 6.15: 処理の有無による候補順位別累積正解率（151 標題）

第 10 候補までの累積正解率では，200 標題の場合で 70.4 %，151 標題の場合で 83.1 %の結果が得られた（図 6.14,6.15）．どの候補順位においても，処理ありの方が良い結果が得られているのが分かる．そして第 10 候補まで結果では，処理を行うことにより 200 標題で 6.0 %，151 標題で 7.5 %正解率を向上させることが出来た．

6.6 おわりに

従来の文字認識過程と異なり，対象文字列からの文字切り出しを前提としない文字認識手法を提案した．そして正解率低下の原因である探索範囲設定の失敗パターンを分析し，先頭探索範囲の拡張処理と，文字パターン辞書に対する上下のストローク幅切除という，前処理を行う事で正解率の向上を試みた．その結果 200 標題の場合で 6.0 %，151 標題の場合で 7.5 %累積正解率を向上させることが出来た．しかし図 6.7(d) のパターンについては，今回改善を行えなかったので検討していく必要がある．さらに他の古文書文献に対しても，同様の実験を行って行きたいと考えている．また更なる正解率の向上のためには，知識ベースの導入が有効であると思われる [22],[10],[7]．候補文字抽出の際や，抽出後の候補順位の入れ替えなどに知識ベースが利用できれば，処理時間の短縮や，正解率の向上が期待できる．

今後は正解率の向上を目指すだけでなく，GUI によるユーザインターフェースを作成し，対話型システムの検討を行いたいと考えている．

第7章

『くずし字解読辞典』文字画像からの筆順抽出の試み

7.1 『くずし字解読辞典』文字画像からの筆順抽出の試み

7.1.1 はじめに

翻刻の支援のために，まず実現が望まれているのがくずし字の検索システムである．この実現には，くずし字の文字認識辞書が欠かせない．このため，HCR プロジェクトでは，児玉幸多編『毛筆版くずし字解読辞典』(東京堂出版)（以後『くずし字辞典』）を出版社の許諾を得てデジタル化した．この電子版『くずし字辞典』を使って，われわれは日本語入力 FEP を使って文字を入力すると，そのくずし字と，さらに類似した文字の画像を表示するソフトウェア (e-Kuzushi) を作成した．しかし，このソフトウェアは文字から翻刻を検索するシステムであるため，利用者が不明な文字の見当がつかないと検索できないという欠点がある．このため，われわれはくずし字から翻刻を検索できるシステムの開発に取り組んでいる．くずし字の入力方式としては，(1) オンライン入力 (タブレットやペン入力などで文字を手書き入力する方式)，(2) オフライン入力 (スキャナなどで文字を画像入力する方式)，の 2 方法が考えられる．オンライン入力では筆順情報が利用可能，オフライン入力ではつづけ字の切り出し処理が必要，といったことから，オンライン入力された文字の認識のほうがより容易であると考えられる．オンライン入力された文字を認識するためには筆順情報を備えた文字認識辞書が必要となるため，われわれは電子版『くずし字辞典』に収容されている各文字画像に筆順情報を付加することにした．最初は筆順情報を得るために，マウスカーソルの座標を表示できる画像処理ソフトウェアでくずし字画像を表示させ，ヒトが座標値を読み取って表計算ソフトウェア上で入力するという作業を行っていた．しかし，この方法は相当の作業量・作業時間を要し，非実用的であることが明らかとなった．このため，筆順抽出のためにいくつかのツールや技法を開発した．本報告は，この過程で得られた知見をまとめたものである．

7.1.2 筆順情報取得支援ソフトウェアの開発

ソフトウェアの概要

筆順情報を容易に抽出できるように，筆順情報取得支援ツールを作成した．これは，利用者が，画面上に表示されるくずし字の画像を見ながら，線の中心点をマウスでクリックしていくことで，文字の筆順情報を得ようというプログラムである．作成したプログラムのユーザインタフェースを図 7.1 に示す．ウインドウ左下の画像ファイ

ルの一覧から，対象となる画像ファイルを選択すると，右側のボックスに画像が表示される．ボックス上でマウスカーソルを移動すると，ボックス上での xy 座標の値がウインドウ上の pos X 及び pos Y に表示される．筆の中心上にあると思われる点にマウスカーソルを移動し，クリックすることで，その点の座標値が記憶される．これを繰り返すことで，クリックした順に座標の系列が取得できる．1 つのストロークのサンプリングが終了したら，Shift キーを押しながらクリックする．これによって（0, 0）というデータが格納され，ストロークの終わりを認識できるようになっている．記憶された座標の系列は，ファイル名とともにテキストファイルに出力できる（図 7.2）．

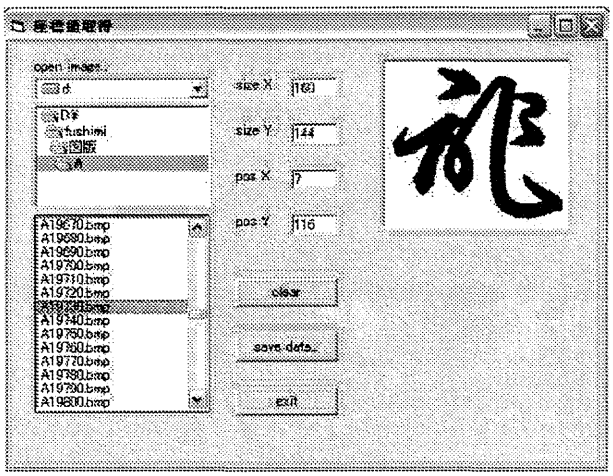


図 7.1: 作成したプログラムのユーザインタフェース

A19730.bmp	
51	7
71	18
51	33
0	0
19	60
13	72
60	50
31	105
41	80

ストロークの区切り

x 座標　　y 座標

図 7.2: 出力されたデータ例

7.1.3 問題点

作成したソフトウェアを利用することにより作業効率は向上したが，やはり手動であるため，辞書の収録文字全ての筆順データを得るためには膨大な時間が必要である．このため，筆順抽出処理の自動化を試みた．

7.2 筆順自動抽出の試み

くずし字の筆順を推定するために，次の方針で処理を行うことにした．

1. まず，くずし字を端点や交点で区切られた文字の部分品に分割する．
2. 次に部分品を接続してストロークを得る．
3. 最後にストロークの順序を推定する．

本稿では処理 1 に焦点を絞って報告する（処理 2，処理 3 の実現法は目下研究中である）．処理 1 のために，文字画像から文字の骨格線を取り出すことにした．文字画像の骨格線を抽出するために，細線化を行う方法が知られている．しかし，くずし字に対して細線化を適用するには問題がある．例えば文字「す」のように，筆がループ状に動く場合，書き方によっては中心にある空白が潰れてなくなってしまう場合がある．このような画像を細線化すると，その部分が一本の線になってしまう，あるいは無くなってしまう（図 7.3 の矢印が示す部分．この細線化は (http://cse.naro.affrc.go.jp/sasaki/slim/slim.html) のプログラムによる）．この問題を解決するために，筆の中心点から骨格線を取り出す方法を試みた (次節で述べる)．結果的にこの方法は成功しなかったため，次々節で，新たに工夫した探索円を用いる方式について述べる．

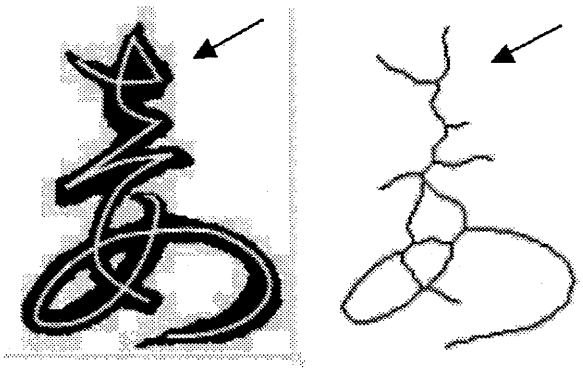


図 7.3: 細線化における問題

7.2.1 中心点から骨格線を取り出す試み

文字画像上で，左から右，上から下，左上から右下，右上から左下，の 4 方向に対し，「白から黒になる点」（A 点）と「黒から白になる点」（B 点）を探し，各方向の A − B 点間の中心点を求め，プロットする（図 7.4）．これによって得られた結果が図 7.5 である．

この結果から骨格線を構成する点を得ようと試みたが，正確な骨格線を推定できるまでには至らなかった．そこで，別の方法を試みた．

7.2.2 細線化せずに部分品に分割する方式

くずし字に対して細線化を行ってしまうと，文字の輪郭の連続性など，筆順を判定するヒントとなる貴重な情報が失われてしまう．細線化を行わずに部分品に分割できれば，その後の処理を行う上で有利となる可能性があるため，そのような方式を考案した．本節ではその方式を述べる．以後，文字画像上で，文字が描かれた黒い部分を文字領域，文字領域と白い紙の部分との境界をエッジと呼ぶ．

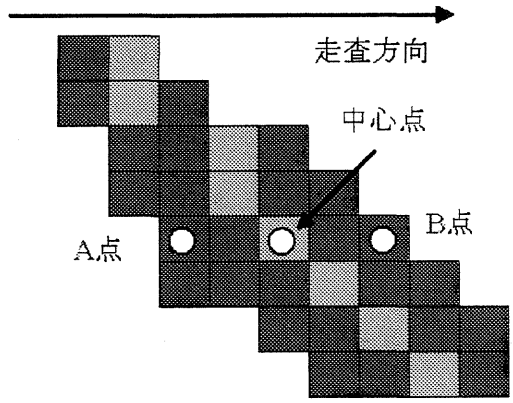


図 7.4: 中心点取得の例（x 軸方向）

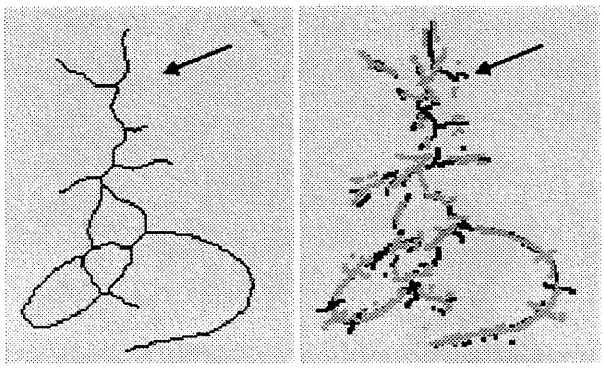


図 7.5: 細線化と中心点取得による方式の比較

方式の概略

ここで提案する方式は，筆の線の中心線上を，半径が可変の円 (探索円と呼ぶ) で辿っていくことにより文字部品に分割するというものである．探索円は，以下のように動かす．まず，文字領域上で探索円を配置する (探索円の初期位置の決定法は次節で述べる)．次に，以下の処理を繰り返す．

- 円内に，必ず線の両側のエッジが入るように，円の半径と中心を調整する．
- 中心点から，両側のエッジの方向を使って筆の走っている方向を求め，探索円を少し移動させる（図 7.6)．このとき，中心とエッジの座標を記録する．

円を移動させていくうちに，両側のエッジがつながってしまえば，端点に到達したとみなす．円内に，注目しているエッジと別のエッジを検出した場合，分岐点に遭遇したとみなす (図 7.7)．この場合は，再帰的にすべての方向を探索する．このようにして全ての領域の探索を行う．

探索円の初期位置の決定

探索円の初期位置は，筆の線の中心付近にある必要がある．これは以下のようにして見つけている．まず画像をスキャンして文字領域の点を１つ見つけ，仮の探索円の中心を置く．探索円の半径を，非常に小さい値から徐々に大きくしていく．円内にエッジを１つ見つけたら，円の中心をエッジと反対方向に動かす．以上の処理を，エッ

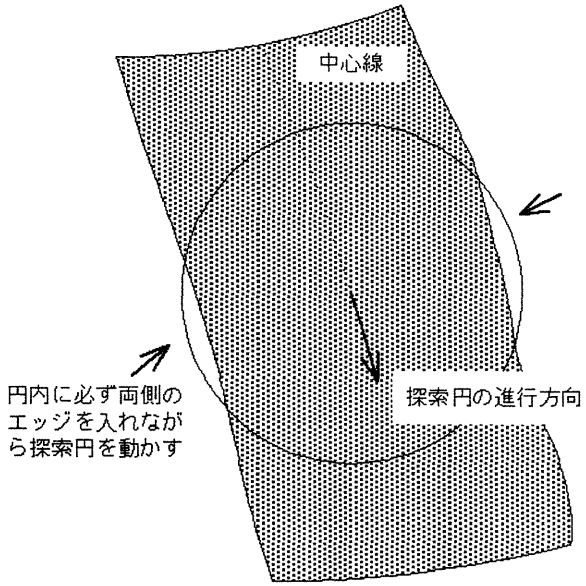


図 7.6: 探索円の動き

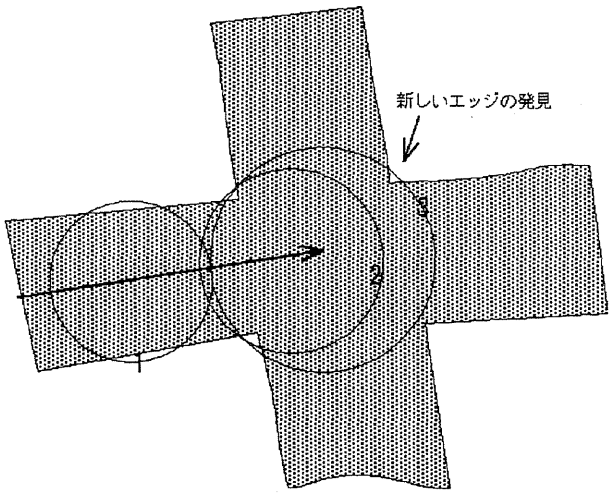


図 7.7: 分岐点の発見

ジを２つ以上見つけるまで繰り返す．この方法によってほぼ満足できる初期位置を得られることがわかった．

文字領域が複数ある場合

上記アルゴリズムでは，一筆書きが可能な画像しか処理できない．「い」のように，複数の文字領域から成る文字に対応するため，次のような方法を採っている．先に述べたように，文字画像をスキャンして解析の開始点を決めるが，決めた後，文字画像上で，その点と繋がっている文字領域を，塗りつぶしアルゴリズムを用いて判定し，マークする．次にマークされていない文字領域から再度解析開始点を決める．これをマークされていない領域がなくなるまで続ける．

実行結果

本アルゴリズムを Java 言語で実装した．解析した結果は，内部的には文字の端点や交点を節点，それをつなぐ部分を辺とする無向グラフとして表現している．図 7.8，図 7.9 に本アルゴリズムで解析した実際のくずし字を示す．わかりやすいように，認識した部分品ごとに異なった色をつけるようにした．また，節点には数字が，辺にはアルファベットが振ってある．かなり正確に部分品に分割できているのがわかる．

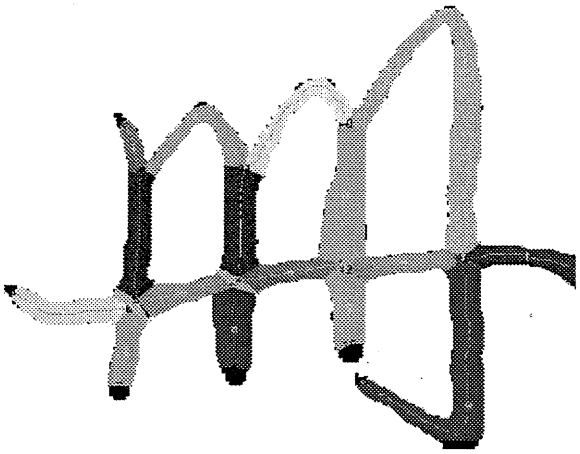


図 7.8: 実行例 1

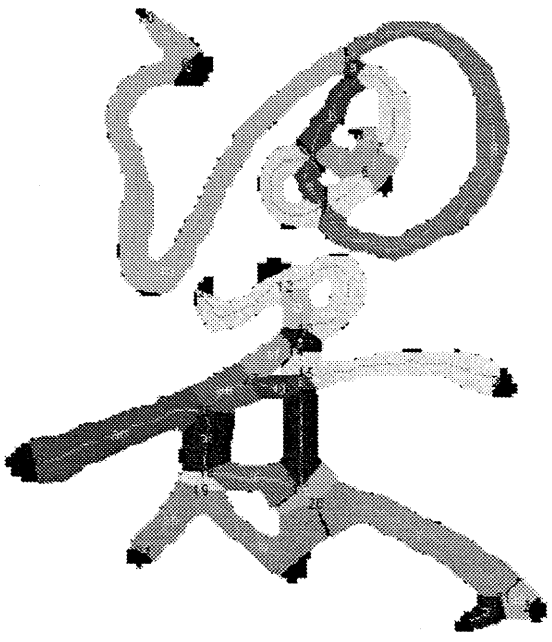


図 7.9: 実行例 2

ただし，適切な結果が得られない場合も存在する．前述の，ループの部分がつぶれてしまっているような場合，本アルゴリズムでは単に太くなっている線だとみなされてしまう．これは，辺の太さの頻度分布や，エッジの滑らかさを調べることで何らかの対処ができる可能性があり，今後の課題の一つである．

7.3 おわりに

本報告では，デジタル化されたくずし字辞書に対して筆順情報を追加するための試みについて述べた．本研究では，まず，文字画像を見ながら手動で筆順を得るための支援ツールの開発を行った．手動では限界があることがわかったため，この処理の自動化を試みた．このためには，まず文字を部分品に分割する必要がある．筆の中心点列から抽出する方法を試みたが，よい結果が得られなかったため，筆の輪郭線も考慮して線分を辿る手法を開発した．この方法により，ひとまず満足する結果が得られた．

今後は筆順を自動判定する方法を研究・開発する予定である．基本的には得られたグラフ上の全ての節点について，そこから伸びている辺のどれとどれとがひとつのストロークで描かれたのかを，何らかの評価値を用いて推定するアルゴリズムになると考えている．また，この技術はオフライン文字認識にも応用することが可能であろう．

第 8 章

知識による翻刻支援

8.1 はじめに

古文書には多くの種類があるが，近世の借金証文類は様式が比較的一定しており，使用されている用語には定型がある．たとえば，「依而如件」「実正也」などの用語は必ずといってよいほど文書のなかに登場する．その他の用語についても，借金証文のなかでよく使われるものがみられる．

借金証文のように使用される用語に定型がみられる種類の文書については，多くの用例を集めてそこから用語に関する知識を抽出し，知識にしたがって翻刻者を支援する方法が考えられる．具体的な方法としては，n-gram を利用することの有効性が予想される．

われわれは古文書証文類を対象に，翻刻時に遭遇する読めない文字（不明文字）の前後文字から n-gram の情報を使って不明文字の正解候補を提示する可能性について検討した．証文類の用例データとするために「伏見屋文書」の全文を翻刻した．さらに，本手法を実装した翻刻支援ユーザインタフェースを作成し，被験者を用いた利用試験を実施し，その結果，システムの有効性を確認することができた．

8.2 n-gram による不明文字候補検索実験

8.2.1 検索手法

n-gram による不明文字の正解候補検索手法は，つぎのとおりである．

検索対象である不明文字を c_i とすると，その前後の文字のつながりは，

$$\cdots c_{i-1}c_ic_{i+1}\cdots$$

と表現される．

一方，用例データから得られる n-gram テーブルはつぎのように定義される．

$$t_{j,1}t_{j,2}\cdots t_{j,n},f_j$$

ここで $t_{j,1}$ は用例中に登場する n 文字のつながりの 1 文字目，$t_{j,2}$ は n 文字のつながりの 2 文字目，f_j はその n 文字のつながりの頻度である．

n-gram テーブルからの不明文字の正解検索は，前方一致の場合と後方一致の場合にわけられる．前方一致は $c_{i-n+1}\cdots c_{i-1}$ と $t_{j,1}\cdots t_{j,n-1}$ のマッチングをとることであり，後方一致は $c_{i+1}\cdots c_{i+n-1}$ と $t_{j,2}\cdots t_{j,n}$ のマッチングをとることになる．

前方一致のケースと後方一致のケースにおける候補文字の確率を総合して，つぎのような第 1 候補文字 $t_{k,n}$ の選択基準を定義する．

前方一致した n-gram の集合を $\{t_{k*}\}$，後方一致した集合を $\{t_{l*}\}$ とすると，

$$\max_{t_{k,n}} F(f_k, f_l) = \max(f_k, f_l; t_{kn} = t_{l1}).$$

以下，$F(f_k, f_l)$ の降順に，$t_{k,n}$ を第 2 候補，第 3 候補…とする．

8.2.2 用例データベース

n-gram による古文書翻刻支援のための用例データとするために，大阪市立大学所蔵の「伏見屋文書」の全文を翻刻した．その結果，用例データ量は約 243,000 文字となった．

8.2.3 不明文字検索実験結果

古文書翻刻中に遭遇する不明文字の正解候補を，用例データから作成した n-gram を用いて検索することの有効性を試験した．「伏見屋文書」全文データから無作為に 10 文書を選択して，それらの日付と署名部分を除く表題と本文部分を試験データとし，残りの文書の全文データから 5-gram までを作成して教師データとした．n-gram の作成は，長尾らの方法 [23] によった．

試験データの全 1,553 文字を 1 文字づつ取りだし，それらを不明文字と仮定して教師データから作成した n-gram をもとに不明文字の正解候補を出した．$n = 2$ から 5 までについて n-gram から不明文字の正解候補を出し，候補文字中の累積正解出現率を第 50 候補まで求めたものが，図 8.1 である．

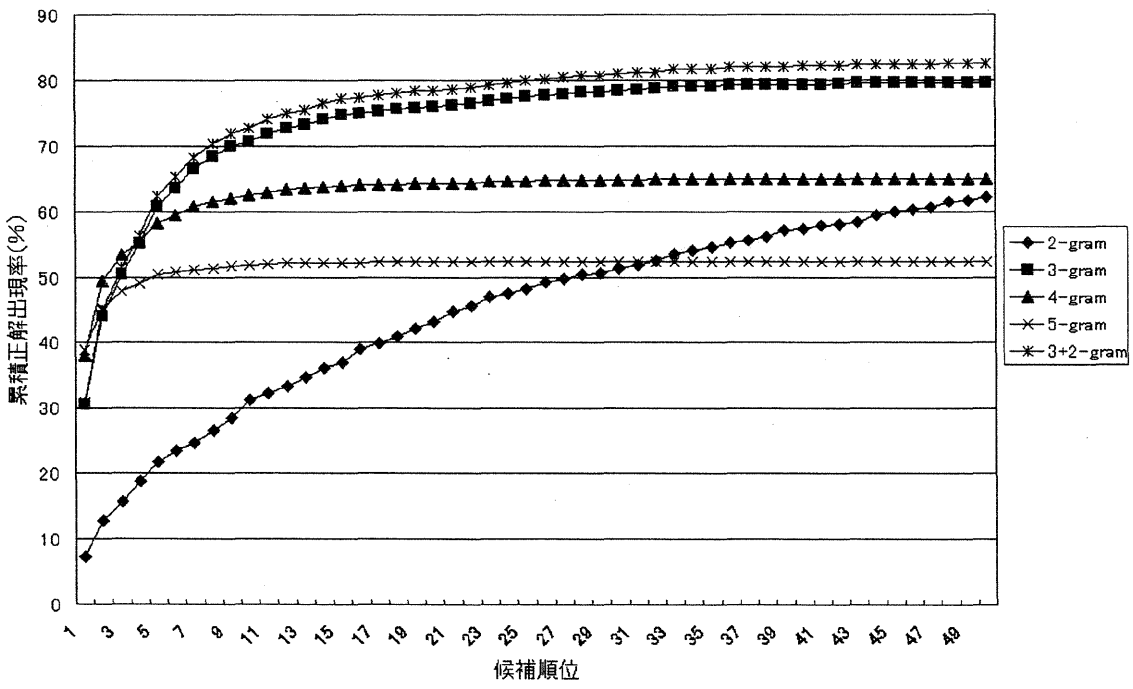


図 8.1: $n = 2$ から 5 までの候補順位別累積正解出現率（第 50 位まで）

図 8.1 によると，$n = 2$ から 5 までの間の累積正解出現率は $n = 3$ で最大となることがわかる．したがって，古文書翻刻支援のためには，用例データの 3-gram を知識として用いることが適当であると考えられる．

2-gram から 5-gram までで候補文字が得られたなかった割合を示したものが，図 8.2 である．3-gram では候補文字が得られなかった割合が 5.8 %であるのに対して，2-gram ではすべての不明文字に対して候補文字が得られた．2-gram は図 8.1 にみられるように正解出現率の点で 3-gram に劣るものの，候補文字を提示する能力におい

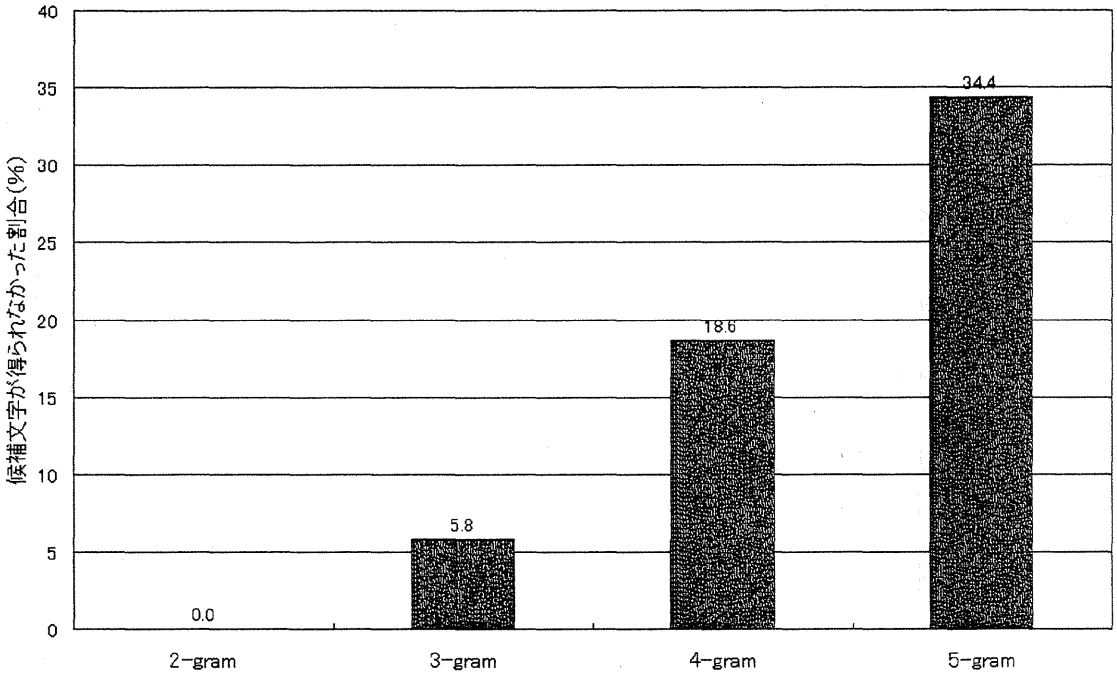


図 8.2: 候補文字が得られなかった割合

ては 3-gram よりも優れている．したがって古文書翻刻支援のためには，3-gram で正解候補を示し得ない不明文字に対しては 2-gram を適用することが有効であると考えられる．実際に 3-gram で正解候補が得られなかった場合に 2-gram を適用する手法（以降 3+2-gram とする）を用いて，おなじ試験をしてみた結果が，図 8.1 中の 3+2-gram のグラフである．

図 8.3 は，3+2-gram で得られた正解候補数の頻度分布である．正解候補数の平均値は 18.47 候補，最頻値は 1 候補，最大値は 286 候補であった．

表 8.1: 第 10 候補までに正解があらわれた累積割合 (3+2-gram)

候補	累積割合（%）
1	30.97
2	44.95
3	51.44
4	56.34
5	62.40
6	65.23
7	68.26
8	70.32
9	71.93
10	72.70

システムとしての実用性を考慮した場合，正解が第 10 候補までに入ることをひとつの目安としうる．表 8.1 は，3+2-gram を用いた場合の第 10 候補までに正解があらわれた累積割合である．第 10 候補までに正解があらわれた割合は，72.70 %であった．また正解があらわれた最高は第 250 候補で，その累積正解出現率は 83.77 %あった．

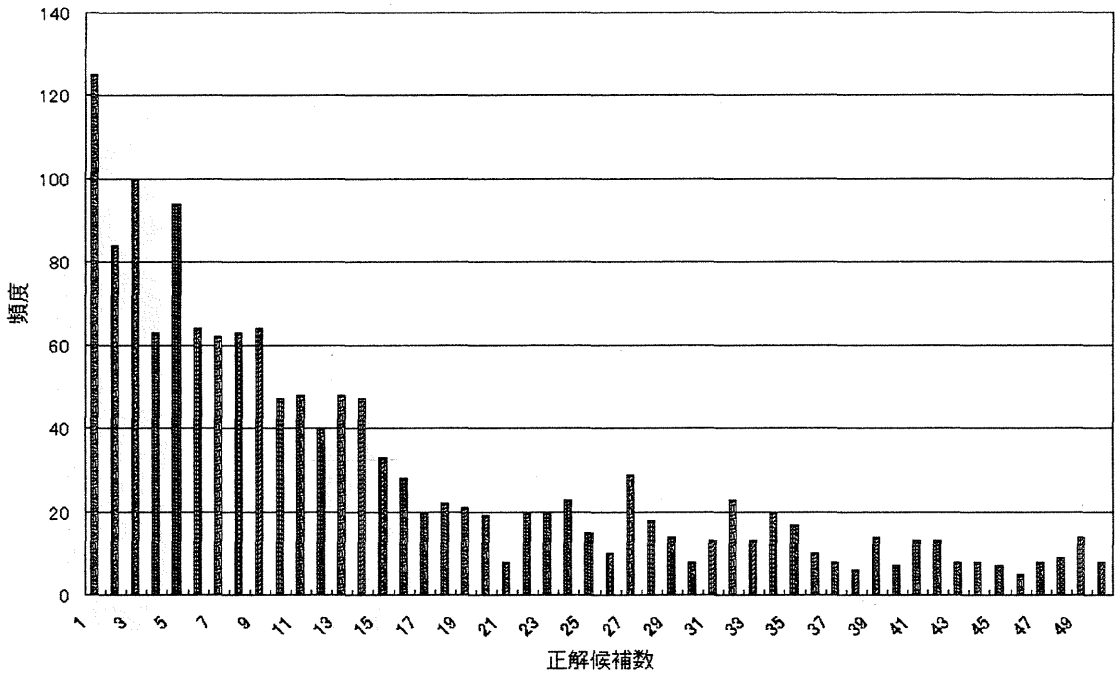


図 8.3: 正解候補数の頻度分布 (3+2-gram，50 候補まで)

8.3 GetAMoji マクロの利用試験

8.3.1 ユーザインタフェースの実装

「伏見屋文書」の全文用例データから 3+2-gram を用いて不明文字の正解候補を提示する機能を持った，翻刻支援のためのユーザインタフェース（GetAMoji マクロ）を試作した．ユーザインタフェースは，Microsoft Word 2000 のマクロ言語である Visual Basic for Application を利用して作成した．Word の操作画面から本手法による GetAMoji マクロを呼び出し，正解候補を Word 入力画面に反映できるようになっている．画面例を図 8.4 に示した．

8.3.2 利用試験

GetAMoji マクロの有効性を試験するために，古文書翻刻経験のない被験者を使って利用試験を実施した．被験者に「伏見屋文書」のなかの 1 文書の紙焼きを示し，その表題と本文部分のみを辞書など参考資料を一切使わずに自分の力で翻刻し，翻刻文を Microsoft Word で入力してもらった．解読できない不明文字は，「□」で入力するよう支持した．その作業が終了した後，Word 上で GetAMoji マクロを起動し，システムから提示された「□」の部分の候補文字をみて，被験者が正解と思った文字を「□」と置換した．システムの教師データからは，翻刻対象文書の用例データを除外した．

作業時間の制限は設けず，被験者が納得いくまで作業してもらった．被験者は 30〜40 歳代の男女 3 名である．被験者はいずれも古文書翻刻の経験はないが，1 名（被験者 A）は入門程度の古文書読解教育を受けたことがある．

被験者ごとの利用試験結果を，マクロ使用前と使用後でまとめたものが表 8.2 である．3 被験者を平均すると，マクロの利用によって正解文字数は 9.3 %増加し，不明文字数は 10.8 %減少したが，不正解文字数も 1.5 %増加し

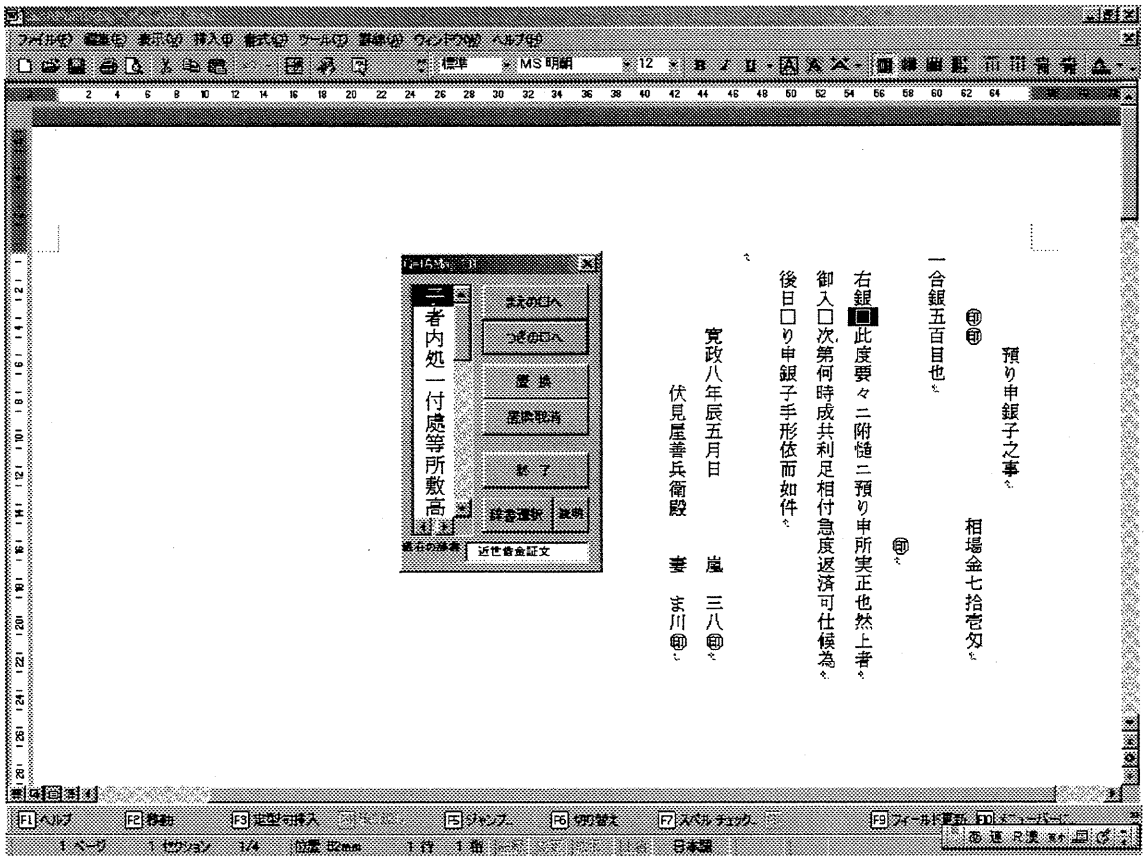


図 8.4: GetAMoji マクロ

表 8.2: GetAMoji マクロ利用試験結果

	被験者 A		被験者 B		被験者 C		平均	
	使用前	使用後	使用前	使用後	使用前	使用後	使用前	使用後
正解文字数	47(69.1)	55(80.9)	23(33.8)	26(38.2)	22(32.4)	30(44.1)	(45.1)	(54.4)**
不正解文字数	5(7.4)	5(7.4)	13(19.1)	12(17.6)	22(32.4)	26(38.2)	(19.6)	(21.1)
不明文字数	16(23.5)	8(11.8)	32(47.1)	30(44.1)	24(35.3)	12(17.6)	(35.3)	(24.5)*

括弧内は%. $*: p < 0.1, **: p < 0.05$.

た．しかし平均の片側 t-検定の結果，正解文字数は有意水準 5 %で，不明文字数は有意水準 10 %でそれぞれ増加しているといえるが，不正解文字数は有意に増加しているとはいえないことがわかった．したがって，GetAMoji マクロは古文書翻刻支援に有効であるといえることが確かめられた．

被験者の誤りについてさらに詳しく分析してみる．GetAMoji 使用前は不明文字で，使用後も不明文字のままであるか不正解文字となった文字のうち，GetAMoji が提示した候補文字のなかの第 1 候補に正解があった文字数は，被験者 A が 1 文字，被験者 B が 6 文字，被験者 C が 3 文字であった．また第 10 候補文字のなかに正解があった文字数は，それぞれ 2 文字，8 文字，5 文字であった．これらの文字は，GetAMoji が上位に正解をあげていたにもかかわらず，被験者に古文書読解知識がないために，正解として認知されなかった誤りである．その 1 例を図 8.5 に示した．1 文字目の「然」について，被験者 A は GetAMoji 使用によって正解を得たが，被験者 B と C は正解が全 11 候補中の第 1 位に示されていたにもかかわらず，この文字を「然」とは翻刻しなかった．3 文字目の「者」も第 1 候補に正解が示されていたが，すべての被験者が「者」とは翻刻しなかった．

図 8.6 の部分は，すべての被験者が末尾の「為」は翻刻できたが，他の部分は文字の切り出しができず，文字数

図 8.5: 被験者が正解を認知できなかった 1 例（然上者）

図 8.6: 被験者が正解を認知できなかった 1 例（急度返済可仕候為）

すら判断できない結果となった.

8.4 おわりに

以上の結果,「伏見屋文書」の全文を対象として，前後の既知文字から 3-gram および 2-gram の情報を使って不明文字を検索する実験により，第 10 候補までで 72.70 %の正解率を得られると推定できた．さらに本手法を実装した GetAMoji マクロの利用試験をおこなったところ，翻刻経験のない初心者が辞書なしで翻刻した結果の正解文字数が有意に増加することがわかり，マクロの有効性が確かめられた．この結果は，辞書を併用した場合や翻刻経験者が使用した場合のさらなる有効性を示唆するものである.

本手法は，不明文字の前後の文字が正しいと仮定して，その情報から不明文字の候補を提示するものである．したがって，前後の文字がそもそも誤っていたり，文字数の推定が誤っていたり，不明文字が連続してしまった場合には，正しい候補文字の提示ができない．本手法の応用として，英文のスペルチェックに対応するような，翻刻済み文字に対する検証システムのようなものも考えられるだろう．また本手法は，証文類という一定の表現が頻出するパターンをとる文字列に対して有効な手法であって，その他の種類の文書対してこの手法がどの程度有効であるかは今後の検討が必要である.

第 9 章

知識と OCR による文字の推定

9.1 はじめに

古文書の翻刻支援に目的を絞った場合，高精度の文字認識は差し当たり絶対に必要な条件とはならない．なぜならば，古文書の完全自動読み取りとは違って，翻刻支援の場合は人間が介在する作業を効率化するような情報をシステムが提供できればよい．つまり，たとえ不完全であっても，人間による推論の助けになる情報を提示することが重要なのである．したがって，翻刻支援システムとして利用価値のあるものにするためには，システムが出力する第 1 候補文字が正解である率を 100% に近いレベルで競うことよりも，たとえば正解が候補文字の上位 20 位に入る割合を 80% 程度にすることが，現段階での目標になる．

この論文では，江戸時代の借金証文類に対象を限定して，翻刻作業中に遭遇する判読不能な文字（不可読文字）を，その前後の文字の n-gram 情報と不可読文字の画像データの OCR 結果から不可読文字の正解を推定する方法を検討し，当手法を古文書翻刻支援システムに応用した場合の有効性を，大量の実データを使って検証する．

江戸時代の借金証文類に対象を限定する理由は，この種の文書には「預り申所実正也」「急度返済可仕候」「仍而如件」といった定型表現が頻出するため，n-gram のような統計情報で不可読文字を推定できる可能性がたかいからである．さらに，借金証文のような江戸時代の公文書は「御家流」という書体で筆記されているため，文字のくずし方にある程度の法則性がみられ，毛筆・くずし字という OCR に不利な条件が緩和される．また，借金証文類の翻刻は江戸時代の経済史研究にとって重要な作業であるにもかかわらず，未翻刻の文書数は，各地の文書館や個人の蔵で眠っているものも含めると，それこそ無数にある．したがって，借金証文類に対象を限定した研究であっても，実用への期待と可能性はたかいといえる．

われわれは，実証性を重視した検証を進めるために，江戸時代の借金証文類 231,161 文字を翻刻して用例データを作成し，それらのうちの 3,509 文字についてくずし字のなかから 1 文字を切り出した文字画像データを作成した．さらに，標準的な古文書文字辞典から 24,244 文字を採字して，その文字画像データと文字データを電子化し，OCR のための学習データにした．これらのデータを使って，n-gram 情報と OCR のそれぞれによる不可読文字の推定と，両者を総合した推定結果を示し，翻刻支援システムにこの手法を適用した場合の性能と有用性について考察して，情報処理学のあらたな適用分野の開拓を試みる．

9.2 n-gram 情報による不可読文字の推定

9.2.1 方法

用例データから作成する n-gram[23] は $n = 2$ と $n = 3$ を併用し，不可読文字の推定に当たって $n = 3$ では候補が得られなかった場合に $n = 2$ の情報を使用する方法を採用した．この方法は，本論文の実験で使用するもの

と同種の古文書データを使って，有効性がすでに検証されている [10]．方法の概略は，以下のとおりである．

推定対象である不可読文字を c_i とすると，その前後の文字のつながりは，

$$\cdots c_{i-1}c_ic_{i+1}\cdots$$

と表現され，一方 n-gram テーブルは，

$$t_{j1}t_{j2}\cdots t_{jn}, f_j$$

と表現される．ここで t_{j1} は用例データ中に登場する n 文字のつながりの 1 文字目，t_{j2} は 2 文字目，f_j はその n 文字のつながりの出現頻度である．

n-gram 情報を使って不可読文字を推定する方法は，文献 [?] では前方一致と後方一致が取られているが，本論文ではそれらに加えて $n = 3$ の場合の中間一致も考慮することにする．すなわち，不可読文字 c_i に対して，

- 前方一致した集合:
$$F_f(c_i) = \{(t_{k3}, f_k)|t_{k1} = c_{i-2}, t_{k2} = c_{i-1}\}$$
- 中間一致した集合:
$$F_m(c_i) = \{(t_{l2}, f_l)|t_{l1} = c_{i-1}, t_{l3} = c_{i+1}\}$$
- 後方一致した集合:
$$F_b(c_i) = \{(t_{m1}, f_m)|t_{m2} = c_{i+1}, t_{m3} = c_{i+2}\}$$

となり，不可読文字 c_i の正解候補の集合 $G(c_i)$ には，前方・中間・後方一致のうち頻度が最大となるつぎのような要素を与える．

$$\begin{aligned} G(c_i) &= \{(t_{**}, f_*)\} \\ &= \{\max_{f_*}(F_f(c_i), F_m(c_i), F_b(c_i)) \\ &\quad |t_{k3} = t_{l2} = t_{m1})\} \end{aligned}$$

n-gram 情報からの推定による正解候補のスコア $NScore$ は，頻度の合計からの比例配分値の逆数であるつぎのような値を与える．

if $f_* > 0$

$NScore(t_{**}) = \Sigma f_*/f_*$

else

$NScore(t_{**}) = 1$

$NScore$ は $(0 < NScore \leq 1)$ の値をとる．ただし，不可読文字が n-gram テーブルに対して前方・中間・後方のいずれにも一致しない場合は，$NScore$ は不定とする．すなわち，

if $\Sigma f_* = 0$

$NScore(t_{**}) = NONE$

計算順序は，まず $n = 3$ で $NScore$ を求め，それが不定になる場合に限って $n = 2$ で同様の操作をおこなう．$NScore$ は小さいほど良好な推定となる．

9.2.2 実験

実験には，「伏見屋文書」の全文翻刻を用いた．「伏見屋文書」は金融・借家・親族関係に関する議定書などからなる総数 1,300 の文書群で，翻刻後の総文字数は 231,161 文字である．そこから，後述する OCR の実験にも用いる 30 文書 3,509 文字の試験データを除いたものを用例データとして n-gram を作成した．すなわち，用例データと試験データは重複しない．

翻刻にあたっては，古文書の文字を MS 明朝フォントが表示する SJIS コードの範囲内でもっとも近い字形を取る文字コードを選択した．したがって，たとえば「返済」と「返濟」がおなじ意味であっても，それぞれ異なる用例として扱われている．

実験では，試験データの 3,509 文字のすべての文字を 1 文字づつ順に取り出して仮想の不可読文字として，正解候補の何番目に正解が出現するかを調べる方法をとった．

9.2.3 結果と考察

提案手法によって，試験データ全体の 79.62% にあたる 2,794 文字について正解候補が得られた．正解候補が得られながら，そのなかに正解が含まれなかった事例は，この試験データ中にはなかった．正解順位の平均値は 5.42 位 ($\sigma = 8.76$)，最頻値は 2 位，最大値は 129 位であった．正解が候補の 1 位となった割合は 8.49%，10 位以内に入った割合は 71.19%，20 位以内では 76.63% であった．一方，正解候補が得られたものの候補数の平均値は 18.10 個 ($\sigma = 22.33$)，最頻値は 1 個，最大値は 290 個であった．

システムとしての実用性を考えると，正解候補として出力される候補数は 20 個程度以下，もし可能ならば 10 個以内であることが望ましいと思われる．あまりにおびただしい数の正解候補を示されても，人間の作業の助けにならないからである．提案手法で得られた正解候補数の平均値は 18.10 個で，20 個以下に収まっている．

しかしながら，試験データの 20.38% にあたる 715 文字について，提案手法では正解候補が得られなかった．すなわちこれらの 715 文字は，その前後の文字列が用例データにマッチしなかった文字である．

正解候補が得られた仮想不可読文字について，平均値で 5.42 位に正解が位置するという結果は，翻刻作業の支援システムとして実用可能な水準であろう．一方で試験データの 20.38%，すなわち平均して 5 文字に 1 文字は，正解候補が出力されないという結果は，翻刻作業支援システムとしての実用化に向けて障害となる．したがって，n-gram 情報になんらかの補助的な情報を加えて，正解順位を向上させると同時に，候補を出力しない例を削減しなければならない．

9.3 OCR による不可読文字の推定

9.3.1 方法

n-gram 情報に加える補助的な情報として，不可読文字の画像情報を与えて，その OCR 結果を加味して総合的な順位を求める方法を試みる．その前にまず，古文書文字の場合に OCR でどの程度の認識率が出るかを検討する．

OCR にはさまざまな文字特徴量の求め方があるが，われわれは日本語手書き文字認識研究で ETL9B データベースに対してたかい認識率を出している改良型方向線素特徴量 [24] をそのまま適用してみることにした．改良型方向線特徴量は，文字を非線形正規化した後に文字の輪郭線を構成する線分の方向の分布を小領域ごとに重み付けをして抽出する方法で，特徴量は 196 次元のベクトルとして得られる．

OCR で高認識率を出すためには，文字認識用辞書をどのように作るかが重要である．われわれは，専門の翻刻者の間で標準的な辞書のひとつになっている『くずし字解読辞典』[5] を選択して，その本編ならびに付録に掲載されている文字画像と，それらに対応する非くずし字・読みなどの情報の文字コードを電子化して文字認識用辞書を作成した．文字画像の電子化は，辞書のページを 400dpi2 値でスキャニングし，1 文字づつを手作業で切り出す方法をとった．このようにして電子化した総文字数は 4,795 字種 24,244 文字である．

『くずし字解読辞典』では，ひとつの文字について 2 種類のくずしのパターンが例示され，その非くずし字と読みが示されている．おなじ文字の異なるくずし文字が複数の場所に掲載されている場合もあるので，得られるサンプル数は文字によって異なるが，ひとつの文字に対するサンプル数は非常に少ない．1 文字あたりのサンプル数

の平均値は 5.06 個 ($\sigma = 4.43$)，最大値は 55 個，最頻値は 2 個である．

『くずし字解読辞典』では，くずし字に対応する非くずし字は活字ではなく手書きであるため，われわれは手書き非くずし字にもっとも近い字形の SJIS コードを与え，SJIS コードに対応する文字がない場合は今昔文字鏡コードを割り振った．その際，たとえば「済」と「濟」がおなじ文字であるといった字形の包摂概念については考慮せず，与えた文字コードが異なっていればそれらは別の文字として取り扱った．『くずし字解読辞典』のうち SJIS コードを割り振ることができたのは，4,053 字種 22,061 文字である．

このようにして，『くずし字解読辞典』から抽出した文字画像について，改良型方向線素特徴量を算出し，文字認識用辞書とした．文字認識は，試験データの文字画像から改良型方向線素特徴量を求め，文字認識用辞書のなかからユークリッド距離が近い順に正解候補を選択し，そのユークリッド距離を認識スコアとする方法をとった．その際，正解候補中におなじ文字コードを持つ候補が複数出現した場合は，それらのうちのユークリッド距離の最小値をもってその文字の認識スコアとした．

9.3.2 実験

古文書文字認識の試験データとして，「伏見屋文書」から 30 文書 (3,509 文字) をランダムに選択し，そのすべての文字を手作業で切り出して文字画像データを作成した．試験データの作成は，作業進行上の制約により，つぎのような手法をとった．

1. 原文書をスキャニング
2. 画像をいったんシートにプリント
3. 専門の翻刻者がマーカーで 1 文字を囲むようにシート上に記入
4. マーク済みシートをスキャニング
5. マーキングされた 1 文字を画像から自動切り出し
6. 2 値化してノイズ除去処理
7. 文字コードとの対応づけ

このようにして作成された試験文字画像データについて，前節の方法によって文字認識を施した．正解候補の算出にあたっては，計算時間の短縮のため，文字認識用辞書データ数の 5% にあたる上位 1,212 文字まで候補を求め，それ以下の順位をとる候補は切り捨てた．

9.3.3 結果と考察

実験の結果，正解が 1,212 位までに入ったものは，試験データ全体の 73.64% にあたる 2,584 文字で，正解順位の平均値は 112.80 位であった．

この結果は，この方式による OCR 単独では古文書翻刻支援のための実用にはほど遠いことを示している．しかしこれは，つぎの理由からじゅうぶんに予想される結果であった．第 1 に，文字認識用辞書の規模が小さい．すなわち，1 文字あたりのサンプル数が少ない．第 2 に，OCR アルゴリズムは既存の日本語手書き文字認識用のものをそのまま適用しているので，古文書文字に対して最適化がされていない．第 3 に，認識方法としてユークリッド距離法というごく単純な方法を用いている．これらはひとつひとつが大きな研究テーマであるので，本論文の課題からは除外する．

本論文では，OCR としては改良の余地を残す方法ではあっても，そこから得られるあらたな情報を有効に活用する方法を探りたい．本手法は，全体としての認識率の点で劣るとはいえ，なかには良好に文字認識ができている試験データもある．図 9.1 は，OCR で得られた正解候補のうち，20 位以内に正解があった例の累積割合である．OCR の結果，正解が 1 位にきた例は，試験データ全体の 6.41% にあたる 225 例あり，上位 10 位に入った例は 19.12% にあたる 671 例，上位 20 位では 25.79% にあたる 905 例あった．数は少ないとはいえ，これら OCR か

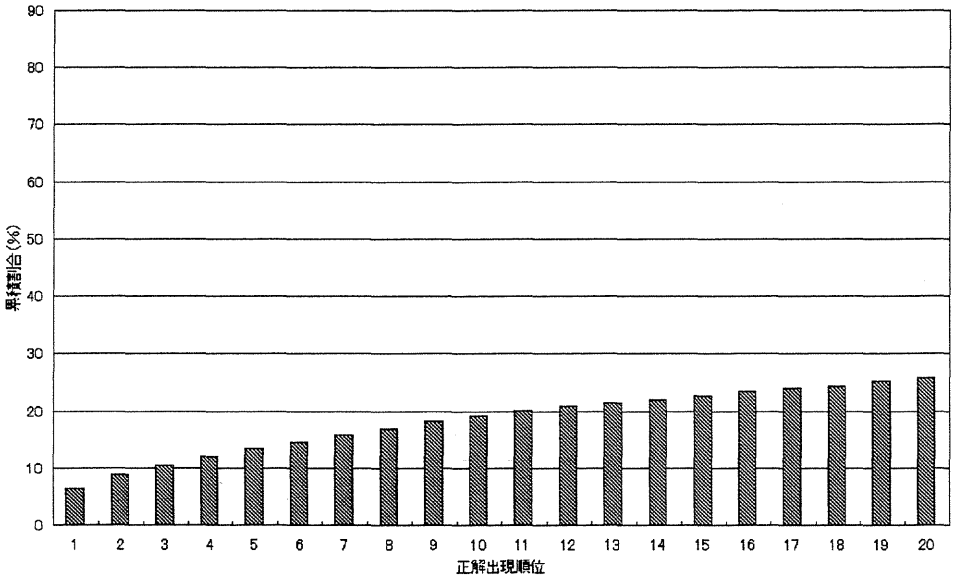


図 9.1: OCR による正解出現順位の累積割合

ら得られた情報を有効に利用し，n-gram 情報による方法と OCR 結果を組み合わせることで，n-gram 情報のみの場合よりも不可読文字の推定結果を向上させうる見込みがある．

9.4 n-gram と OCR の併用方法の考察

n-gram 情報と OCR 結果を併用した総合スコア ($TScore$) を，つぎのように設定する．

if $NScore(t_{**}) \neq NONE$

$TScore(t_{**}) = NScore(t_{**}) * OScore1(t_{**})$

else

$TScore(t_{**}) = OScore2(t_{**})$

現状では OCR の信頼性が低いため，$TScore$ の算出にあたっては OCR 結果のなかでとりわけたかいスコアを出した結果のみを選択して使用することが妥当である．すなわち，試験データと学習データの文字特徴量のユークリッド距離を $ED(t_{**})$ とすると，

if ranking of $t_{**} < Threshold1$

$OScore1(t_{**}) = ED(t_{**})$

else

$OScore1(t_{**}) = NONE$

if ranking of $t_{**} < Threshold2$

$OScore2(t_{**}) = ED(t_{**})$

else

$OScore2(t_{**}) = NONE$

と定式化され，$TScore(t_{**})$ の昇順で t_{**} を正解候補とする．ただし，$OScore1$ がすべて $NONE$ となる場合は，$NScore$ をもって $TScore$ に替える．

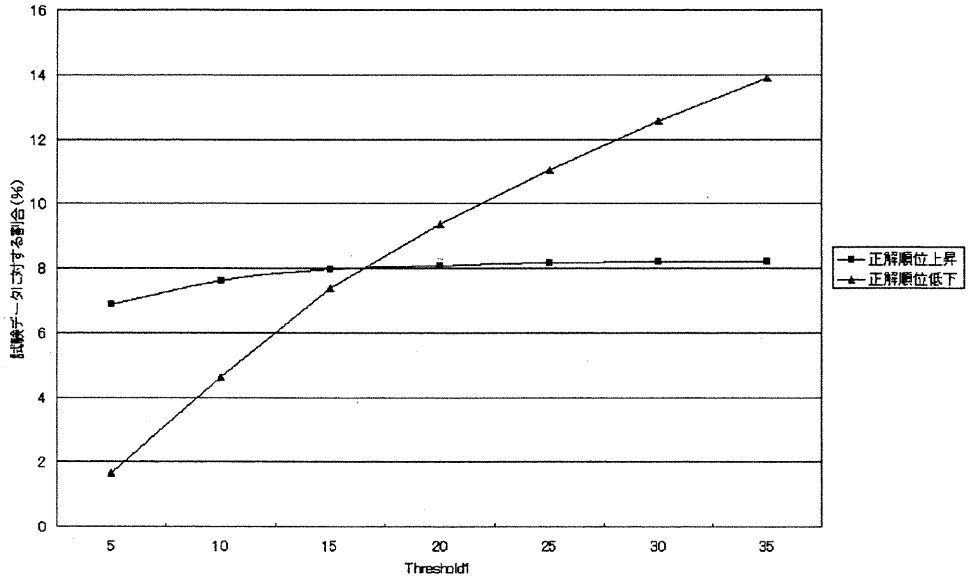


図 9.2: OCR 結果を加味することによる正解順位の変化（n-gram 情報がある場合）

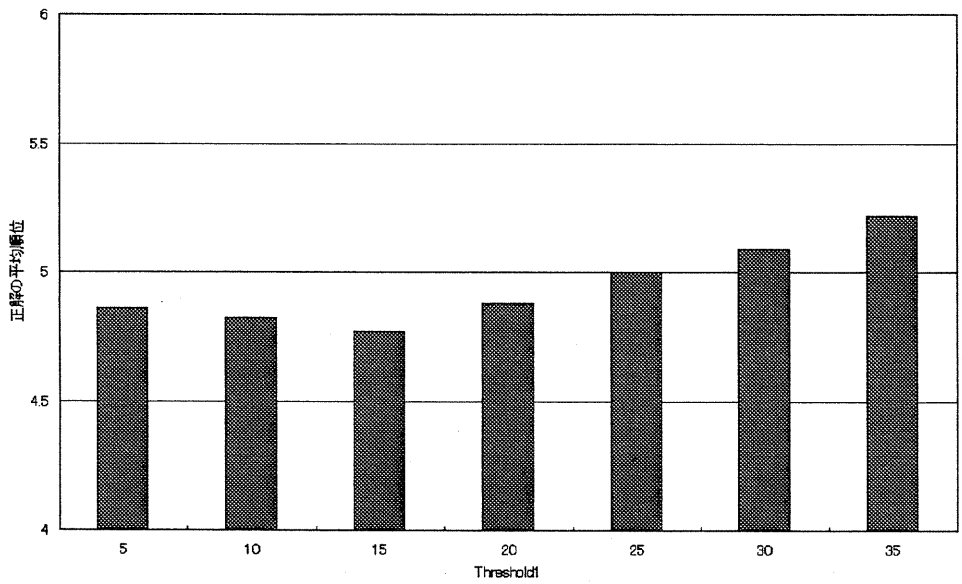


図 9.3: OCR 結果を加味することによる正解の平均順位（n-gram 情報がある場合）

この操作はすなわち，n-gram 情報からの推定結果が上位にあっても OCR 結果が悪い場合はスコアを下げ，前者の結果がさほど上位でなくとも，後者の結果がとくに良ければスコアを上げることになる．また，n-gram 情報から正解候補が得られない場合は，OCR 結果のみから正解候補を出す．

ここで問題になるのは，OCR 結果の正解候補数のしきい値 $Threshold1$ と $Threshold2$ をどのレベルにするかである．

まず，試験データのなかで n-gram 情報から正解候補が得られた 2,794 文字について，OCR 結果の併用を検討する．図 9.2 は，OCR 結果のしきい値 $Threshold1$ を変化させた場合に，n-gram 情報のみの場合と比較して $TScore$ において正解の順位が上昇するか低下するかをみたものである．$Threshold1$ を増加させると，n-gram 情報のみの場合よりも正解順位が上昇する例が漸増するが，$Threshold1 = 15$ 付近で上昇数は頭打ちになる．OCR

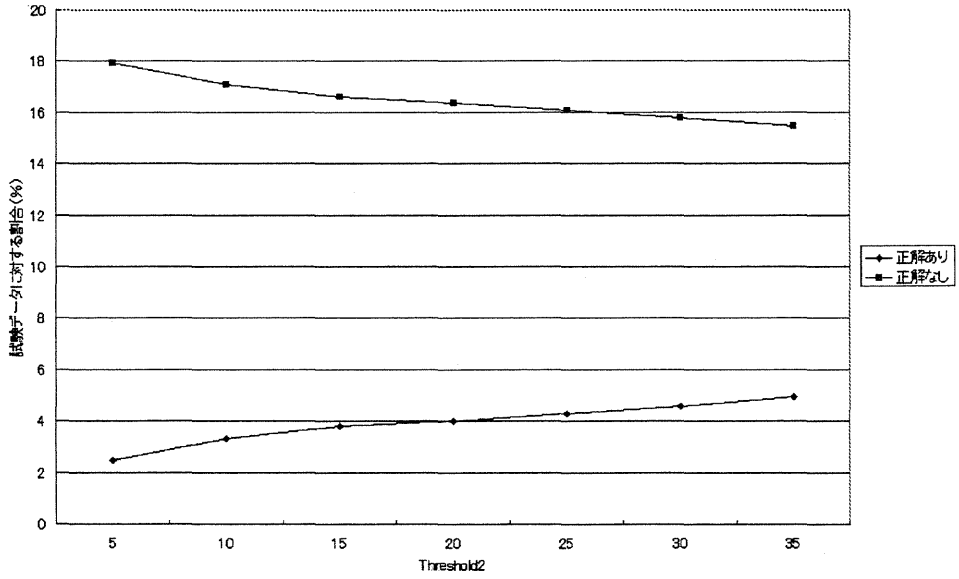


図 9.4: OCR 結果中の正解の有無 (n-gram 情報がない場合)

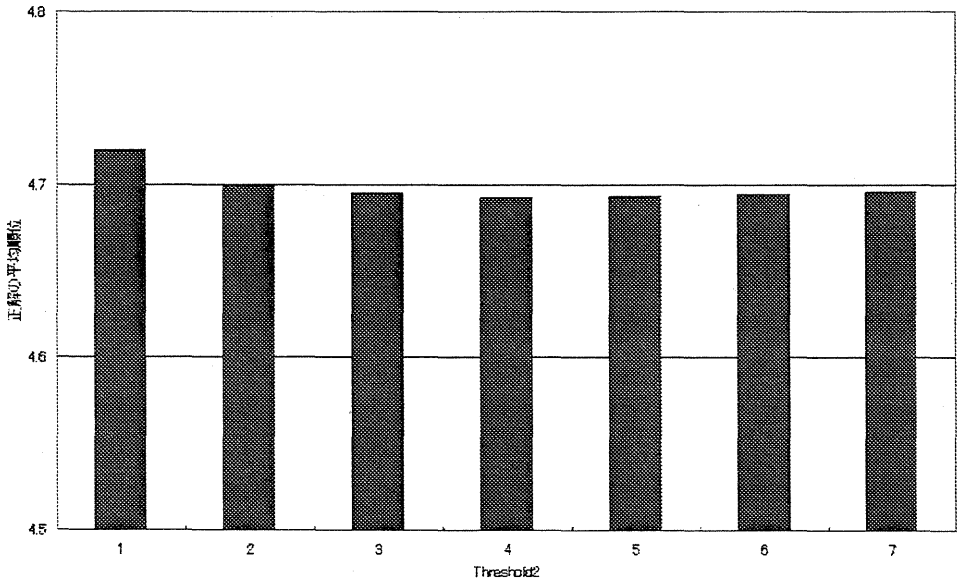


図 9.5: 正解の平均順位 ($Threshold1 = 15$)

結果を併用すると n-gram のみの場合よりも正解順位が低下する例は，$Threshold1$ を増加させるにしたがって上昇する．これらは，OCR の信頼性が低いために，しきい値の操作によって正解ではない候補がノイズとなって，上位に位置するようになるためである．

図 9.2 からは，$Threshold1$ を小さい値にするほど，正解順位の上昇に貢献することになる．しかしながら，図 9.2 では，$Threshold1$ の操作によってどの程度順位が上昇・下降するかについては考慮されていない．

図 9.3 は，n-gram 情報がある場合について，$Threshold1$ の変化による正解の平均順位をみたものである．図 9.3 によると，$Threshold1 = 15$ 付近で平均順位が 4.77 位となることがわかる．この付近の整数値を調べたところ，実際に $Threshold1 = 15$ で平均順位がもっともたかくなる．したがって，n-gram 情報による候補が得られた試験データについては，$Threshold1$ をこの値に設定するのが妥当であると考えられる．

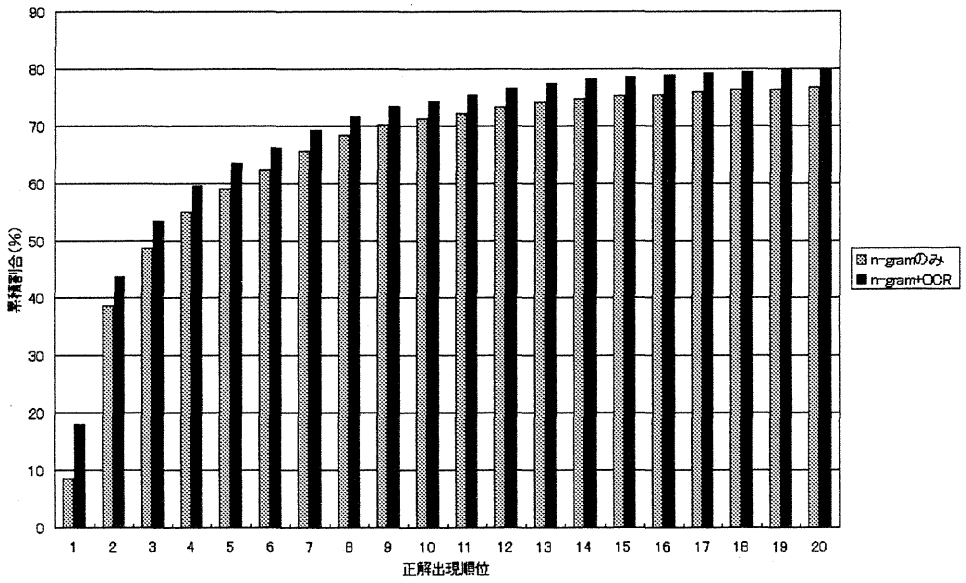


図 9.6: n-gram と OCR の併用による不可読文字の正解順位の分布 ($Threshold1 = 15, Threshold2 = 4$)

つぎに n-gram 情報により候補が得られなかった 715 文字について検討する．n-gram 情報で候補が得られない場合，OCR 結果のみを情報として用いる．図 9.4 は，これらの 715 文字について OCR にかけてみた結果である．当然のことながら，$Threshold2$ の値を大きくするにしたがって候補中に正解が出現する割合はたかくなるが，候補数は $Threshold2$ 個得られることになる．候補中に正解が出現する率と比較して正解が出現しない率のほうがたかいため，$Threshold2$ の値を大きくすることは，正解を含まない候補をむやみに多く出力する結果を招く．したがって，これら 715 文字についても，$Threshold2$ の妥当な水準を決定する必要がある．

$Threshold2$ の妥当な水準の決定方法として，$TScore(t_{**})$ を基準にした場合の，候補中の正解順位の平均値を最小化する方法を採用することにする．図 9.5 は，その結果である．$Threshold2 = 4$ で，正解の平均順位が最小の 4.69 位 ($\sigma = 7.54$) となった．

図 9.6 は，n-gram 情報のみの場合と OCR 結果を併用した場合とで，不可読文字が候補中の何位にあらわれるかを比較したものである．両者を併用した場合，正解順位の最頻値は 2 位，最大値は 129 位となった．これらのしきい値で正解が 1 位となる文字数は，全体の 17.98% にあたる 631 文字，正解が 10 位以内に入る文字数は，全体の 74.35% にあたる 2,609 文字，20 位以内だと全体の 79.77% にあたる 2,799 文字である．この結果は，n-gram 情報のみの場合に正解の平均順位が 5.42 位，1 位が 8.49%，10 位以内が 71.19%，20 位以内が 76.63% であったのと比較すると，不可読文字の推定性能が上昇していることを示している．とくに，正解が候補の 1 位となる割合について，OCR 結果を併用することの効果が顕著である．同時にこれらのしきい値では，全体の 18.07% で正解を含まない 4 個の候補を出力することになる．

9.5 おわりに

本論文では，古文書の翻刻作業中に遭遇する不可読文字について，前後の文字の n-gram 情報と不可読文字画像の OCR 結果を併用して正解候補を求める手法を提案し，250,000 文字を超える古文書文字データと 27,000 文字を超える古文書文字画像データを電子化して手法の検証をおこなった．提案手法により 3,509 文字の試験データの 81.93% について，正解の平均順位が 4.69 位，20 位以内に正解が得られる割合が 79.77% という結果が得られた．これらの結果は，提案手法を古文書翻刻支援システムに実装した場合の有効性を示唆するものである．

ただし，本論文で問題にした不可読文字にはいくつかのタイプがあり，提案手法では対応できないものもある．

たとえば，n-gram 情報では不可読文字の前後の文字はただしく翻刻されていることが前提になる．前後の文字がそもそも誤って読まれていたり，不可読文字が連続する場合には，提案手法ではよい精度は得られない．また，OCR では背景ノイズが少なく，「にじみ」や「かすれ」の少ない文字画像が必要である．古文書の文字では，紙の虫食いなどによって文字の一部が欠けてしまっていて，OCR がそもそも不可能な例も多い．

これらの限界はあるものの，古文書の全自動読み取りではなく，あくまで人間の作業を支援するシステムのための方法として，提案手法がある程度有効である可能性を示すことができたのではないかと考えられる．本論文では OCR の認識方法については，ごく初歩的な方法をとった．今後 OCR を古文書のために最適化することにより，不可読文字の推定精度がさらに向上することが，じゅうぶんに期待できる．

第 10 章

電子化古文書文字辞典

10.1 はじめに

現在のところ，古文書くずし字辞典類のなかで電子化されたものはない．古文書翻刻のさいに使用される標準的な辞書を電子化し，検索の利便性をたかめることができたならば，翻刻作業の向上が見込まれる．電子化を考えるならば，現在もっともよく使われている辞書を対象にするすることが理想である．翻刻者がよく使用している辞書のひとつに，東京堂出版『毛筆版くずし字解読辞典』[5]（以後『くずし字辞典』）がある．この辞書は，文字の第一ストロークの方向を検索キーにしている点が，ほかの辞書にみられない特徴となっている．不明な文字を調べるさいに，第 1 画の方向から探索することができる．しかしながら，この辞書を実際に使ってみると，求める文字にたどりつくにはそれなりの時間がかかり，検索漏れもおこる．電子辞書化して検索の方法を工夫することで，知りたい文字にたどりつくまでの時間を短縮し，検索漏れをすくなくすることができるだろう．

辞書を電子化することによって，紙の辞書では到底できない検索方法をとることができる．それは，ある文字に類似した文字を一覧的に検索することである．類似文字検索を実現するさいに鍵となるのは，文字の特徴量と文字間の類似度の設定方法である．日本語手書き文字認識技術で使用されている手法を応用することで，くずし字の特徴量と類似度を求めることができる．

以上のようなアイデアのもとに，『くずし字辞典』を電子化し，類似文字検索機能について検討して，電子古文書文字辞典を実装した．

10.2 辞書の電子化

電子化の対象としたのは，『くずし字辞典』のなかの付録部分をのぞく章に掲載された 23,707 文字である．ここには漢字，かな文字のほかに，「申上候」「より」などの複数の語からなる用例も 1 文字として含まれている．毛筆で書写された文字の画像を 400dpi2 値画像でスキャナ取り込みした．同時に，文字画像に対応する文字のフォントを Windows 内蔵のフォントで割り当て，内蔵フォントにない文字については今昔文字鏡フォントでカバーした．また複数文字からなる用例をのぞくすべての文字について，今昔文字鏡番号を付与した．読みかたの情報は最大で 9 種類となった．作成した文字情報の一部を図 10.1 に示した．

『くずし字辞典』の特徴は，第 1 ストロークの方向によって文字を分類している点にある．すなわち第 1 画を，①下に向かって連続する点で起筆する「縦点」，②右に向かって連続する点で起筆する「横点」，③右上から左下へ斜めに伸ばす棒で起筆する「斜棒」，④上から下へ伸ばす棒で起筆する「縦棒」，⑤左から右へ伸ばす棒で起筆する「横棒」の 5 種類に分けて，その種類ごとに文字が掲載されている．この第 1 ストローク情報を用いることで，検

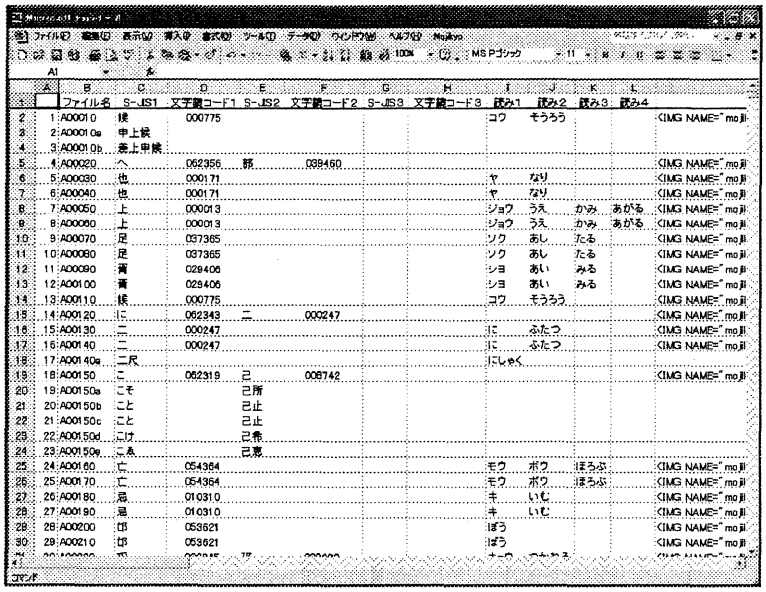


図 10.1: 電子辞書の文字情報

索の精度を向上させることができる．

10.3 類似文字検索手法

文字の特徴量算出方法として，われわれは孫らによる改良型方向線素特徴量 [24] を採用した．同特徴量は，日本語手書き文字認識研究用データベースとして定評のある ETL9B データベースを対象とした実験で，多くの実績とたかい認識性能を示している．特徴量算出の概要を示す．はじめに前処理として，スムージング，輪郭線抽出，正規化をおこなう．スムージングは，文字の局所形状の変化をなめらかにしてノイズを軽減するためのものである．輪郭線抽出によって，文字の外形を取り出す．輪郭線抽出ではなく細線化をおこなうと，文字がつぶれていた場合に文字の形状情報が失われてしまう．その点で，細線化よりも輪郭線抽出のほうが，毛筆の特徴抽出においても優れている．正規化は，津雲による非線形正規化 [25] を採用している．津雲の正規化法は，ストローク間隔の逆数を正規化関数とするもので，ストローク間の間隔をある程度均一化できる．

文字特徴量は，以下の手順で算出する．

1. 輪郭線の線素化
2. 方向線素特徴量の算出
3. 外側加重による方向線素特徴量の補正
4. 方向線素ベクトルの算出

輪郭線の線素化は，輪郭線上の黒画素を方向づける作業である．輪郭線に対して 3×3 のマスクを用いて，線素の方向を縦，横，+45 度，-45 度のいずれかに分類する．ただし図 10.2 のような場合は，たとえば (a) では縦と+45 度の 2 方向に線素があると判断する．

方向線素特徴量の算出方法は，図 10.4 に示した．64×64 ドットからなる文字画像領域を 8×8 ドット単位に分割する．隣接する 4 単位をまとめて 16×16 ドットの小領域とし，縦と横の両方向に，それぞれの半分づつをオーバーラップさせてとっていく．小領域は全部で 7×7 の 49 個えられる．

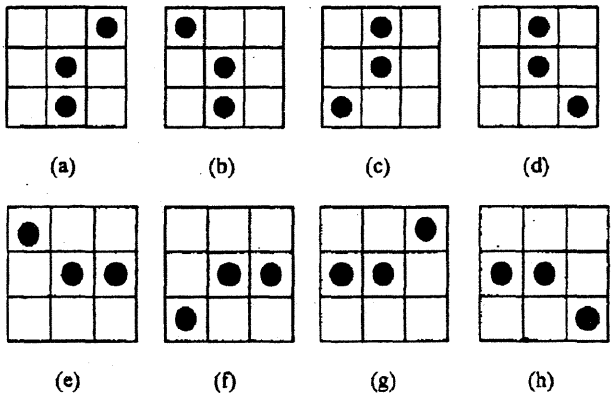


図 10.2: 二つの方向をもつ線素（文献 [24] より引用）

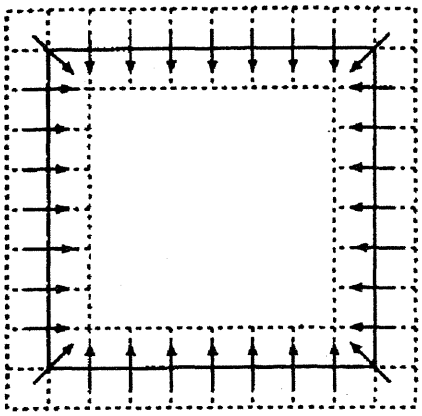


図 10.3: 方向線素特徴量に対する外側加重（文献 [24] より引用）

外側加重による方向線素特徴量の補正は，図 10.3 に示した．文字画像領域の外側に 16×16 個ドットからなる 32 個の仮想小領域を設けて，他の小領域と同様に方向線素特徴量を求め，それぞれ対応する周辺部の小領域の特徴量に加算する．ただし，文字画像領域の 4 隅の小領域には，4 隅を中心とする 16×16 ドットの仮想小領域と，それに半分づつ重複する隣接の仮想小領域の特徴量を加算する．このような外側加重による方向線素特徴量の補正によって，比較的つぶれのすくない文字周辺部の特徴をより有効に利用することができる．

方向線素ベクトルの算出は，つぎの手順でおこなう．各小領域を図??の下図に示すような 4 つの部分に分割し，部分領域にそれぞれ重み 4,3,2,1 を対応させる．各小領域の方向線素特徴量を，つぎの 4 次元のベクトル

$$(x_1, x_2, x_3, x_4)$$

で定義する．ただし，

$$x_i = 4x_{1i} + 3x_{2i} + 2x_{3i} + x_{4i} (i = 1 \sim 4)$$

ここで添字 i はそれぞれ，縦，横，+45 度，-45 度の方向線素を意味する．$x_{1i}, x_{2i}, x_{3i}, x_{4i}$ はそれぞれ中心から外側に向けて 4 つの各部分領域での方向線素 i の個数をあらわす．

したがって 1 文字の方向線素特徴量は，49 個の小領域の方向線素特徴量をならべたもので，次元数は 196 となる．

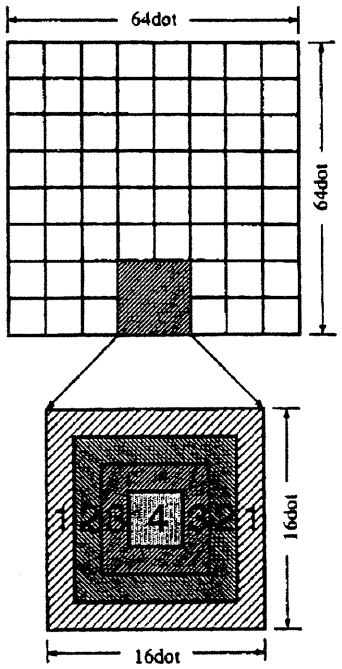


図 10.4: 小領域分割と重み付け（文献 [24] より引用）

各文字間の類似度は，196 次元の方向線素ベクトルのユークリッド距離を使って求めた．文字の類似度は，第 1 ストロークがおなじ文字間についてのみ計算した．そのほうが，第 1 ストローク情報による候補の絞り込みができ，すべての文字間を対象にするよりも検索精度の向上が見込まれるからである．

10.4 電子古文書文字辞典の実装

『毛筆版くずし字解読辞典』に収録された 23,703 文字の画像とテキスト情報，そしておなじ第 1 ストロークをもつすべての文字間の類似度情報をもった電子古文書文字辞典を実装した．実装には，Microsoft 社の Visual Basic 5.0 を使用した．この電子古文書文字辞典は，Windows 環境で稼働する．

現在のところ，検索の入口は文字コードのみとなっている．調べたい文字を ATOK あるいは IME などの日本語入力 FEP を用いて入力し，検索ボタンを押すと（図 10.5），それに該当するすべての文字画像と対応する文字コードが一覧となって表示される（図 10.6）．文字画像にカーソルを合わせると，その読みがちいさなウィンドウに表示される．一度に表示される候補文字は，最大 5 文字になっている．スライドバーを操作することで，右のほうに隠れたほかの候補文字をみることができる．

類似文字を検索したいときは，その文字画像をダブルクリックすればよい．たとえば図 10.6 の右から 2 つめの「預」と似た字を調べたいときは，その画像をダブルクリックすると別のウィンドウが開き，そこに類似文字が表示される（図 10.7）．類似文字の一覧は，検索対象文字とおなじ第 1 ストロークをもつ文字のなかで，196 次元方向線素ベクトルのユークリッド距離がちいさい順に最大第 10 候補まで表示される．類似文字検索はいずれのウィンドウからも可能で，前に開いたウィンドウから直接類似文字検索することもできる．

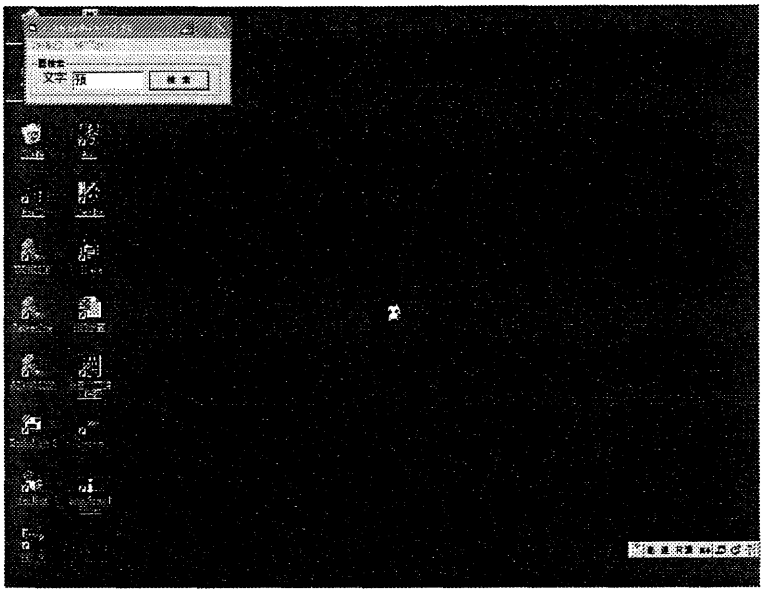


図 10.5: eKuzushi 検索画面 1（「預」を入力して検索）

図 10.6: eKuzushi 検索画面 2（「預」の検索結果）

10.5 おわりに

以上のように，われわれは日本語手書き文字認識技術で使用されている文字特徴量にストローク情報を加味した類似文字検索機能をもった，電子古文書文字辞典を開発した．この電子辞典は，現在のところワープロ的に入力された文字が最初の入口となっているため，わからない文字が何であるかのおおよその検討を利用者がつけなくてはならない．将来的には，①タブレットなどで手書き入力された文字からの検索，②スキャナなどで画像入力された文字からの検索，③より深い階層のストローク情報からの検索機能をもたせるべく，鋭意研究を進めているところである．

図 10.7: eKuzushi 検索画面 3（画面 2 の右から 2 つめの文字に類似した文字の検索結果）

第 11 章

HCR プロジェクトの中間評価

石谷　康人　（(株) 東芝　研究開発センター）
(2002 年 9 月記)

11.1　はじめに

筆者は評価者という立場で，設立当初からＨＣＲプロジェクトに参加しており，プロジェクトの進め方や個別研究方針などを決定する全体会議に参加している．本稿では，全体会議の際になされた成果報告やディスカッションと，これまでにリリースされている研究成果報告書に基づいて，99 年度から 01 年度までの本プロジェクトの成果について評価する．

11.2　プロジェクトの成果

本プロジェクトはこれまでに前例の無い「古文書の自動電子化」を研究対象としている．このため従来の研究成果を参考にできず，手さぐり状態で研究が開始されている．そのため，当面の研究方略が以下のように設定され，それに基づいて全体プロジェクトがいくつかの個別研究に分けられた．

研究方略：

1. 書体の安定した公文書で歴史的な価値の高いものを対象とする．
2. 文字認識のための辞書構築を進めるために，標準文字データベースを作成する．
3. 古文書の読解に関する専門知識を整理し，システム化する．
4. 人間と機械の作業分担を明確化し，両者を円滑につなぐ知的ユーザインタフェースを構築する．

そして，各々の研究において将来的な方向性の見極めを目的としたいくつかの実験や試作がなされ，次の成果が得られている．

研究成果：

(a) 「宗門改帳」から年齢表記文字，単位表記文字，親族関係表記文字など合計 243,000 文字を収集し，データベース HCD1 を作成した．
(b) 「伏見屋文書」から借金証文文字行 600 行，借金証文標題文字 4,933 文字を収集し，データベース HCD2-3 を作成した．
(c) HCD1 を対象として文字認識実験を行い，最高で 99.06 %の性能を実現した．
(d) 「伏見屋文書」から標題を 78.1 %の精度で自動抽出するレイアウト解析技術を開発した．
(e) n-gram 言語モデルを導入した古文書翻刻支援システムを試作した．
(f) 電子くずし字辞典のプロトタイプを開発した．

11.3 プロジェクトの評価

上述したようにＨＣＲプロジェクトではまず研究の方向性の見極めが必要であったことから，分担研究において具体的な成果目標を事前に設定することが困難であった．このため各研究成果がそれぞれの目標に対してどの程度達成されているか評価することは難しい状況となっている．そこで，3 年間の見極めにより現時点で明らかになっていることに基づいて「古文書電子化支援システム」として一つの仮説を立て，現状がそれに対してどのような位置付けにあるか評価することにする．

上記 (b) のデータベース HCD2-3 を作成する際に，すべての工程を手作業で行うとコストが膨大となってしまうことから，次に示す人間と機械の協調によるデータベース作成工程が実現された．

HCD2-3 作成作業：

作業１　古文書をスキャナで画像化し，紙にプリントする．
作業２　プリントされた文書に対し，手作業で文字をマーキングする．
作業３　マーク済み文書をスキャナで画像化する．
作業４　自動文字抽出エンジンにより，マーキングされた文字を切り出す．
作業５　切り出された文字パターンを文字認識辞書と照合し，照合結果を修正することによりデータベースを作成する．

上記 (c) によると，古文書から文字パターンが切り出されていれば，文字種を限定した場合には個別文字認識技術により高精度な電子化を実現できることが分かっている．一方，報告書によれば，古文書にはつづけ字やくずし字が多く，現時点ではそれらを高精度に電子化する認識アルゴリズムが実現されてない．したがって，上述したデータベース作成作業のように機械が不得意とする機能を人間が肩代わりすることにより，文書電子化をすべて手作業で行うケースや，大量の誤りを出力するシステムにより自動電子化するケースより効率良く高精度な電子データを生成することが可能となる．

そこで，以下に示す機能を持つ「古文書電子化支援システム」を考えることができる．

古文書電子化支援システム：

機能１　古文書をスキャナにより画像化し，ディスプレイ画面に表示する．
機能２　オペレータが画面上でオンライン情報入力装置を用いて手作業により文字をマーキングする．
機能３　自動文字抽出エンジンにより，マーキングされた文字を抽出する．
機能４　文字認識エンジンにより，抽出された文字パターンをコード化する．
機能５　n-gram 言語モデルにより文字認識誤り個所を推定し，変換候補を提示する．
機能６　オペレータが手作業で文字認識誤りを修正する．
機能７　くずし字などオペレータが独力で修正入力できない文字に対しては，オペレータが当該文字パターンもしくはオンライン手書き入力により電子くずし字辞典を検索して正しい文字コードを入力する．
機能８　切り出された文字パターンの配置関係に対してレイアウト解析を適用したり，認識結果として得られたコード情報に対してキーワード照合を適用したりすることにより，標題，日付，差出人，受取人などの文書論理構造を抽出する．
機能９　文字認識結果と論理構造解析結果を統合して構造化文書（ＸＭＬ文書）を作成し，文書データベースに格納する．

これまでに得られている研究成果により，このような支援システム構築の見通しを得ることができたことは，当初の研究方略の設定と個別研究の分担化が正しかったと見なすことができよう．しかし，それぞれの研究はこのような目的のもとで連携して実施された訳ではないので性能面で不明なことが多い．今後は，具体的な支援シス

テム構築という目標を設定して研究の分担化と連携・統合を行っていく必要があろう．この場合，必ずしも上述した支援システムを構築する必要は無い．

11.4 今後の課題

筆者は，電機メーカにおいて新聞，論文，雑誌，名刺，表形式文書（帳票），オフィス文書，書籍などを対象としたドキュメントリーダ（文書読取りシステム）を開発・製品化しており，多様な業種やユーザ層に対して製品を提供してきた．これらの製品は，それぞれのユーザに合った利用形態のもとでユーザによって定められた性能仕様に従って運用されている．以下では，このような経験に基づいて，ＨＣＲプロジェクトで想定している「古文書電子化支援システム」が準拠すべき項目を設定し，それぞれの項目においてプロジェクトの今後の課題を列挙する．

応用目的　電子化された古文書に対してどのような応用目的でどのような成果を上げるのか明確にする必要がある．特定種別のデータに対して統計処理を行う場合には，タグ付きデータの生成が必要となるであろう．

対象文書　ＯＣＲ技術の開発は対象文書の幾何的性質や内容により大きく左右されるので技術開発時には対象文書を用いることが望ましい．開発サンプルが最終目的文書と大きく異なる場合には開発が遠回りになる．

電子化量と電子化期間　限られた期間で膨大な量の文書電子化を行う場合には，電子化作業において最もコストがかかる部分に集中して技術開発を行わなければ目標を達成することはできない．

電子化作業の体制（リソース）　電子化作業を行うオペレータ層（専門レベルの特定）を早急に決定する必要がある．オペレータの専門性によって開発すべき技術項目が異なる可能性がある．

電子化作業工程　限られた期間とリソースで電子化作業を行いながら目標を達成するためには，作業工程において人間と機械の分担を適切に行う必要がある．上述した項目を決定した後，速やかに全体作業工程とシステム構成を見積もるべきであろう．

限られた予算，開発リソース，開発期間などの制約のもとで，開発目標を達成して成果を上げることは難しい課題である．筆者のこれまでの経験では，上述した項目において詳細が決定しなければ，開発すべき技術項目とその内容を明確化することはできなかった．さらに，これらの項目がクリアされているかどうかは，開発プロジェクトの成果が目的に沿ったものであるかどうかを判断する評価基準にもなった．したがってＨＣＲプロジェクトでも有意義な成果を出すためには研究の土台となる上記項目を早急に明確化する必要があろう．現状ではプロジェクトは初期段階にあるので評価項目を上述した範囲に限定しているが，プロジェクトが進行するにしたがって技術開発に関する評価項目を増やし，それぞれの評価内容を具体化していく予定である．

第 12 章

発表文献

- 平成 11 年度発表分
  - 尾崎浩司, 柴山守, 荒木義彦: 古文書レイアウト画像のピラミッド型抽象化と標題の自動抽出, 平成 11 年電気関係学会関西支部連合大会発表論文, 1999.
  - 尾崎浩司, 柴山守, 荒木義彦: 古文書画像のレイアウト認識と標題抽出, 京都大学大型計算機センター第 64 回研究セミナー報告, 2000.
  - 山田奨治, 加藤寧, 川口洋, 原正一郎, 石谷康人, 柴山守, 笠谷和比古, 小島正美, 梅田三千雄, 山本和彦: 古文書翻刻支援システム開発プロジェクト報告（1）ープロジェクト概要ー, 情報処理学会研究報告, Vol.2000, No.8, pp.1-8, 2000.
  - 和泉勇治, 加藤寧, 根元義章, 山田奨治, 柴山守, 川口洋: ニューラルネットワークを用いた古文書個別文字認識に関する一検討, 情報処理学会研究報告, Vol.2000, No.8, pp.9-15, 2000.
- 平成 12 年度発表分
  - 尾崎浩司, 柴山守, 荒木義彦 : 古文書画像のレイアウト認識と標題抽出, 情報処理学会研究報告, Vol.2000, No.67, pp.47-54, 2000.
  - 尾崎浩司, 柴山守, 荒木義彦 : 古文書画像のレイアウト認識とラベリング法による標題抽出, 平成 12 年電気関係学会関西支部連合大会発表論文, 2000.
  - 尾崎浩司, 柴山守, 荒木義彦, 山田奨治 : 古文書画像の標題文字セグメンテーション, 人文科学とコンピュータシンポジウム論文集, 情報処理学会シンポジウムシリーズ, Vol.2000, No.17, pp.279-286, 2000.
  - 柴山守 : 証文類古文書標題の文字認識辞書構築とその利用についてー正規化の問題点と文字認識プロセスの検討ー, 京都大学大型計算機センター第 67 回研究セミナー報告, pp.70-79, 2001.
  - 橋本智広, 横田宏, 梅田三千雄 : 自己想起型ニューラルネットによる古文書文字認識, 平成 12 年度電気関係学会関西支部連合大会, 2000.
  - 山田奨治, 柴山守 : n-gram による古文書証文類翻刻支援の検討, 人文科学とコンピュータシンポジウム論文集, 情報処理学会シンポジウムシリーズ, Vol.2000, No.17, pp.185-192, 2000.
  - 海老澤規之, 和泉勇治, 加藤寧, 根元義章 : 非線形正規化を応用した学習パターンの自動生成, 2001 年電子情報通信学会総合大会論文集, D-12-12, pp.179, 2001.
- 平成 13 年度発表分
  - 山田奨治, 加藤寧, 並木美太郎, 川口洋, 原正一郎, 石谷康人, 柴山守, 笠谷和比古, 小島正美, 梅田三千雄, 山本和彦: 古文書翻刻支援システム (HCR) 開発プロジェクト報告 (2), 情報処理学会研究報告, Vol.2001, No.51, pp.9-16, 2001.5.
  - 篠原早苗, 和泉勇治, 加藤寧, 根元義章 : SVM による手書き類似文字認識の学習データ選択と認識精度に関する一考察, 2001 年電子情報通信学会ソサイエティ大会, D-12-8, p183, 2001.
  - Ishitani, Y.: Model-based information extraction method tolerant of OCR errors for document

images, Proceedings of Sixth International Conference on Document Analysis and Recognition, pp.908-915, 2001.
- 石谷康人：データ駆動型処理と概念駆動型処理の相互作用による文書画像レイアウト解析，情報処理学会論文誌，Vol.42, No.11, pp.2711-2723, 2001.
- 橋本智広，梅田三千雄：天保郷帳における石高表記文字の個別認識，情報処理学会研究報告，2002.

• 平成 14 年度発表分
- 山田奨治, 和泉勇治, 加藤寧, 柴山守: 類似文字検索機能をそなえた電子くずし字辞典の開発, 情報処理学会研究報告, Vol.2002, No.52, pp.43-50, 2002.5.
- 山田奨治, 柴山守: 古文書を対象にした文字認識の研究, 情報処理, Vol.43, No.9, pp.950-955, 2002.9.
- 梅田三千雄, 橋本智広: 認識処理を援用した文字切り出しによる古文書キャラクタスポッティング, 電気学会論文誌, Vol.122, No.11, pp.1876-1884, 2002.
- 川口洋: 『江戸時代における人口分析システム（DANJURO ver.2.0)』の構築・運用・利用, 帝塚山大学学術論集, No.9, pp.1-27, 2002.12.
- 近藤博人, 松本隆一, 柴山守, 山田奨治, 荒木義彦: 文字切出しを前提としない古文書標題認識, 情報処理学会研究報告, Vol.2003, No.5, pp.1-8, 2003.1.
- 篠原早苗, 和泉勇治, 加藤寧, 根元義章: SVM を用いた手書き文字認識における学習データ選択と認識精度に関する一考察, 電子情報通信学会技術研究報告, Vol.102, No.708 PRMU2002-256, pp.81-86, 2003.
- 安倍広多, 中塚麻記子, 柴山守: 『くずし字解読辞典』文字画像からの筆順抽出の試み, 大阪市立大学学術情報総合センター紀要, Vol.4, pp.19-23, 2003.3.

• 平成 15 年度発表分
- 山田奨治, 柴山守: n-gram と OCR による定型表現がある古文書の文字の推定, 情報処理学会研究報告, Vol.2003, No.59, pp.17-24, 2003.
- 和泉勇治, 海老澤規之, 加藤寧, 根本義章: 非線形正規化を応用した学習パターン生成による手書き文字認識, 電子情報通信学会論文誌, Vol.J86-D-II, No.10, pp.1391-1399, 2003.

参考文献

[1] 山田奨治：高次局所自己相関特徴による古文書かな文字認識, 情報処理学会研究報告, Vol. 95, No. 14, pp. 21–30 (1995).

[2] 山田奨治：変体かなの認識実験とその応用, 人文学と情報処理, No. 18, pp. 71–75 (1998).

[3] 日置慎治, 上原邦彦, 川口洋：年齢を表記した古文書文字の認識－「宗門改帳」古文書画像データベースを用いた実験－, 情報処理学会研究報告, Vol. 98, No. 11, pp. 35–42 (1998).

[4] 挑戦　古文書ＯＣＲ, 人文学と情報処理, No. 18 (1998).

[5] 児玉幸多編：毛筆版くずし字解読辞典, 東京堂出版, 東京 (1999).

[6] 児玉幸多編：くずし字用例辞典　普及版, 東京堂出版, 東京 (1993).

[7] 尾崎浩司, 柴山守, 荒木義彦, 山田奨治：古文書画像の標題文字セグメンテーション, 人文科学とコンピュータシンポジウム論文集, 情報処理学会シンポジウムシリーズ, Vol. 2000, No. 17, pp. 279–286 (2000).

[8] 柴山守：証文類古文書標題の文字認識辞書構築とその利用について－正規化の問題点と文字認識プロセスの検討－, 京都大学大型計算機センター第 67 回研究セミナー報告, pp. 70–79 (2001).

[9] 橋本智広, 横田宏, 梅田三千雄：自己想起型ニューラルネットによる古文書文字認識, 平成 12 年度電気関係学会関西支部連合大会 (2000).

[10] 山田奨治, 柴山守：n-gram による古文書証文類翻刻支援の検討, 人文科学とコンピュータシンポジウム論文集, 情報処理学会シンポジウムシリーズ, Vol. 2000, No. 17, pp. 185–192 (2000).

[11] 柴山守：古文書の文字切出しを考える, No. 18, pp. 57–63 (1998).

[12] 馬場口登, 塚本正義, 相原恒博：手書き日本文字列からの文字切り出しの基本的考察, 電子通信学会論文誌, Vol. J68-D, No. 12 (1985).

[13] 馬場口登, 塚本正義, 相原恒博：認識処理の導入による手書き文字切出しの一改良, 電子通信学会論文誌, Vol. J68-D, No. 11 (1986).

[14] 尾崎浩司, 柴山守, 荒木義彦：古文書画像のレイアウト認識と標題抽出, 情報処理学会研究報告, Vol. 2000, No. 67, pp. 47–54 (2000).

[15] Rumelhart, M. J. E., D. E. and group, reseqrch P.: *Parallel Distribute Processing, 1, 2*, MIT Press, Cambridge, MA (1986).

[16] Kohonen, T.: *Self-organization and Associate Memory (2nd Edition)*, Spring-verlag, 199–202 pp. (1989).

[17] Powell, M. J. D.: *Radial basis function for multivariable interpolation: a review,IMA Conference on Algorithms for the Approximation of Functions ans Data*, RMCS, Shrivenham, 143–167 pp. (1985).

[18] Sun, A. M., N. and Nemoto, Y.: A Handwritten Character Recognition System by Using Imaproved Directional Element Feature and Subspace Method, Vol. J78-D-II, No. 6, pp. 922–930 (1995).

[19] 日置慎治, 上原邦彦, 川口洋：「宗門改帳」に記録された年齢表記の認識, 人文学と情報処理, No. 18, pp. 64–70 (1998).

[20] 加藤寧, 安倍正人, 根元義章：改良型マハラノビス距離を用いた高精度な手書き文字認識, 信学論 (D-II), Vol. J79-D-II, No. 1, pp. 45–52 (1996).

[21] 井野英文, 猿田和樹, 加藤寧, 根元義章：ストローク情報に基づく手書き郵便宛名の切り出しに関する一手法, 情報処理学会論文誌, Vol. 38, No. 2, pp. 280–288.

[22] 笠谷和比古：古文書における文字認識, 人文学と情報処理, No. 18, pp. 13–18 (1998).

[23] Nagao, M. and Mori, S.: A New Method of N-gram Statistics for Large Number n and Automatic Extraction of Words and Phrases from Large Text Data of Japanese, *COLIN 94 : The 15th International Conference on Computational Linguistics : Proceedings*, pp. 611–615 (1994).

[24] 孫寧, 安部正人, 根元義章：改良型方向線素特徴量および部分空間法を用いた高精度な手書き文字認識システム, 電子情報通信学会論文誌, Vol. J78-D-II, No. 6.

[25] 津雲淳：手書き漢字認識における非線形正規化処理, 昭和 62 年度電子情報通信学会情報・システム部門全国大会, p. 68 (1987).

第II部

付録編

第 13 章

知識による翻刻支援システム GetAMoji マクロ利用マニュアル

13.1 はじめに

13.1.1 概要

GetAMoji マクロは，n-gram という文字の統計情報を使って，古文書翻刻中に遭遇する不明文字（ゲタ文字）の正解候補を提示する機能を持つ，Microsoft Word のためのマクロである．証文類などの定型的な文書の翻刻支援にとくに有効である．

このマクロは，複製，改変，再配布自由であるが，改変した場合はその事実をコード中に明記すること．このマクロは，利用者の責任において利用すること．マクロを使用することによって利用者に損害が生じても，作成者は一切責任を負わない．

GetAMoji マクロは，HCR プロジェクトホームページ http://www.nichibun.ac.jp/~shoji/hcr/からダウンロードできる．

13.1.2 マクロの構成

GAM 辞書作成　GetAMoji 辞書を作成するマクロ
GetAMoji　　マクロ本体

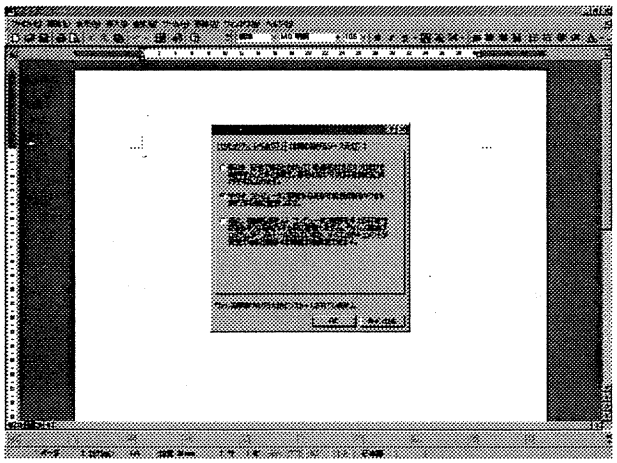
13.1.3 注意事項

Microsoft Windows 98 の Word 2000 で確認済み．Mac の Word では（たぶん）正常に動かない．
マクロで取り扱える n-gram のエントリ数は 10 万エントリまで．それを越えると正常に動作しない．

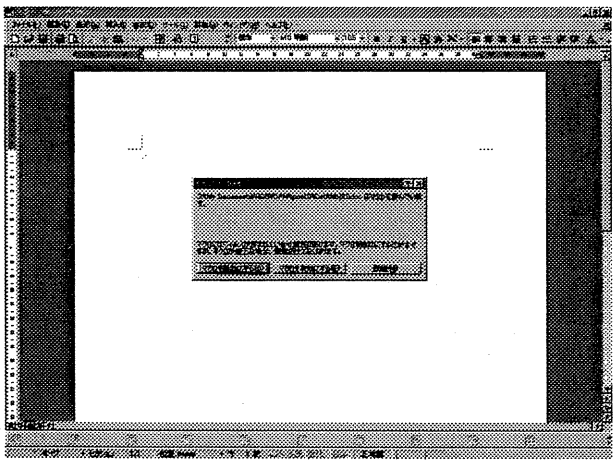
13.2 GetAMoji マクロの利用方法

13.2.1 マクロの登録

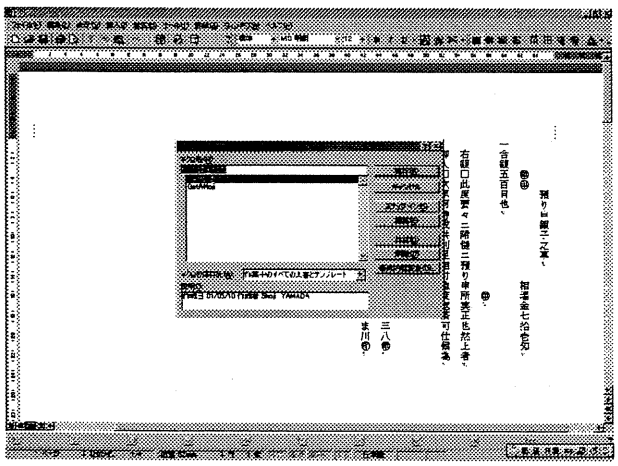
GetAMoji マクロを，あなたの Word 環境で利用できるようにするための手続きである．



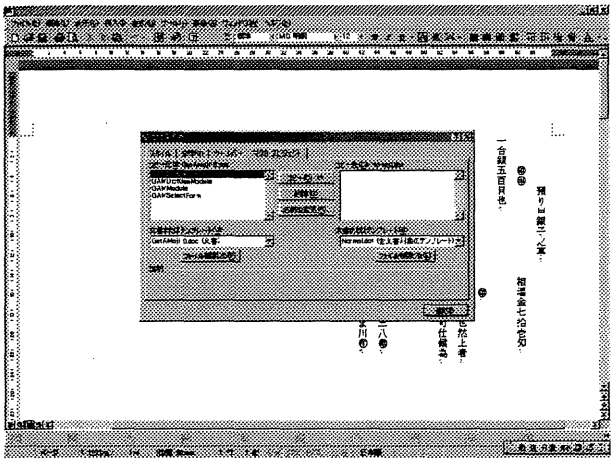
Microsoft Word を起動する．その際，メニューの「ツール」→「マクロ」→「セキュリティ」を選択し，セキュリティレベルを「中」にしておく．



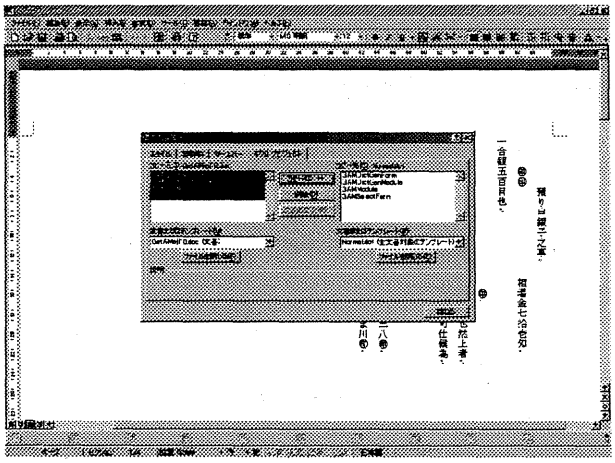
Word に GetAMoji10.doc を読み込む．マクロに関するダイアログが表示されたら，「マクロを有効にする」を選択する．



GetAMoji10.doc が表示されている状態で，メニューから「ツール」→「マクロ」→「マクロ」を選択し，「マクロ」ウィンドウから「構成内容の変更」をクリックする．



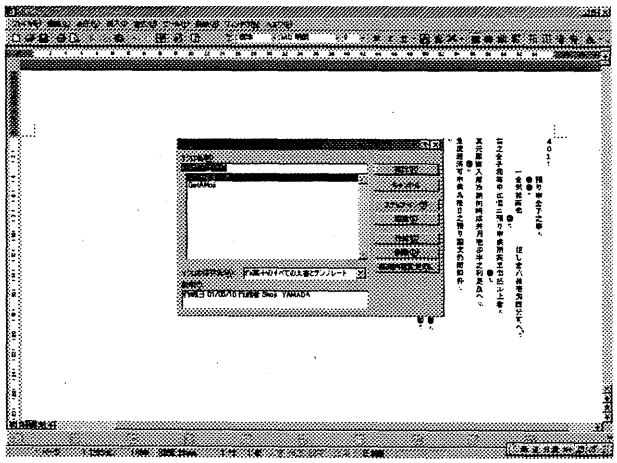
左側の「コピー元」に表示されている 4 つのモジュールを順に選択して「コピー」をクリックし，すべて右側の「コピー元」にコピーする．



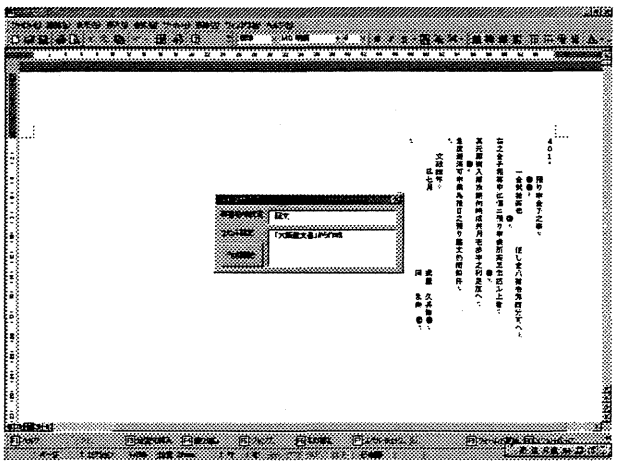
このようになったら，「閉じる」をクリックする．

13.2.2 GAM 辞書作成マクロの利用方法

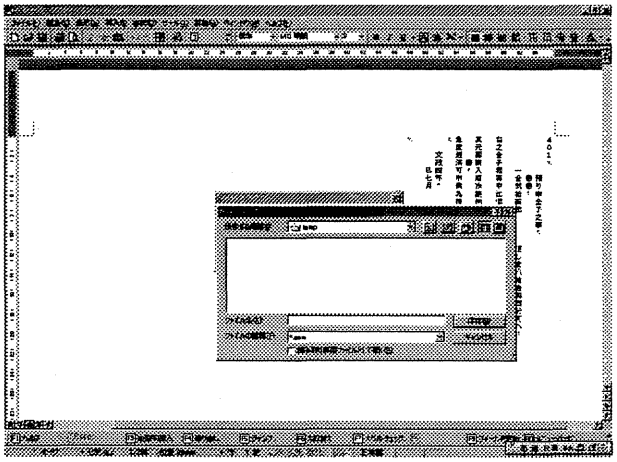
GAM 辞書作成マクロは，あなたの翻刻文から GetAMoji で利用する GAM 辞書を作成するマクロである．



GAM 辞書を作成したい翻刻文を Word に読み込み，メニューから「ツール」→「マクロ」→「マクロ」を選択する．「GAM 辞書作成」を選択して「実行」をクリックする．



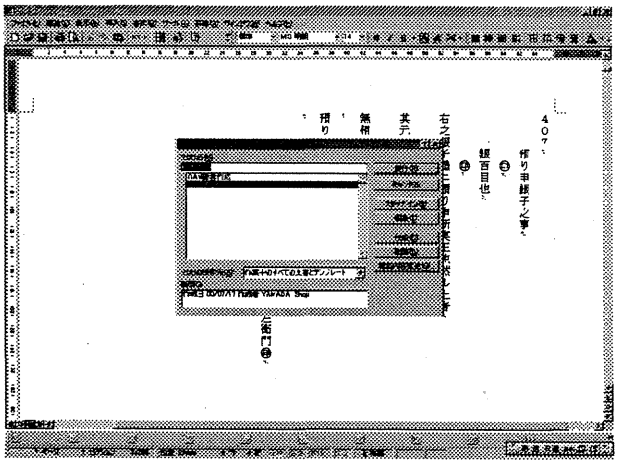
「辞書名称設定」欄に辞書を識別できる名称を，「コメント設定」欄に元データに関するメモを記入して，「作成開始」をクリックする．



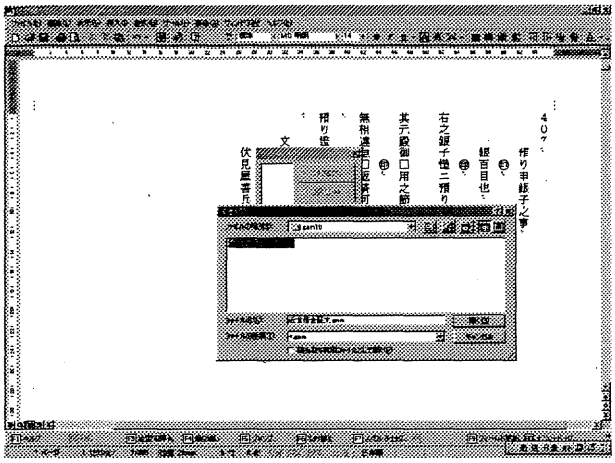
辞書の保存先のディレクトリとファイル名を設定する．設定を完了すると辞書作成がはじまる．処理にはとてもながい時間がかかるので，終了するまでひたすら待つ．

13.2.3 GetAMoji マクロの利用方法

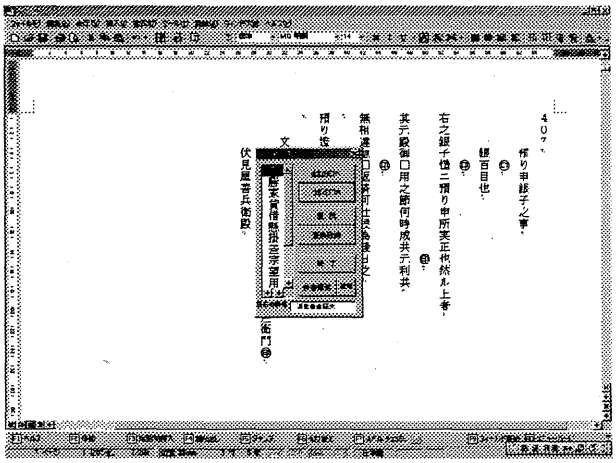
翻刻中の文書のなかの不明文字をあらかじめ「□」にしておく．



翻刻中の文書を Word で表示した状態で，メニューから「ツール」→「マクロ」→「マクロ」を選択し，「マクロ名」から「GetAMoji」を選択して「実行」をクリックする．



「辞書を開く」ダイアログが表示されたら，翻刻中の文書と同じような用語が使われている文書から作成したGAM 辞書を開く．



「つぎの□へ」「まえの□へ」をクリックすると，「□」を検索して正解候補文字を表示する．候補文字のなかに正解とおぼしき文字があれば，それを選択してダブルクリックするか「置換」をクリックすると「□」と置き換わる．まえの置換を取り消したいときは「置換取消」を，GAM 辞書を入れ替えたいときあ「辞書選択」をクリックする．

13.3 効果的な使い方

翻刻対象にあったGAM 辞書があればそれに越したことはないが，なかなかそうはいかないだろう．そこであなたが翻刻したい古文書を，不明文字を□にしたままとりあえず最後まで入力しておく．そしてその翻刻文からGAM 辞書を作成する．作成した辞書を使ってGetAMoji を起動すれば，□に正しい候補文字を出してくれる可能性がたかまる．